Machine Learning in Space Weather

Mandar Chandorkar

To cite this version:

Mandar Chandorkar. Machine Learning in Space Weather. Machine Learning [cs.LG]. Université of Eindhoven, 2019. English. tel-02430788

HAL Id: tel-02430788
https://hal.inria.fr/tel-02430788

Submitted on 7 Jan 2020

Technische Universiteit
Eindhoven
University of Technology

Department of Applied Physics

Machine Learning in Space Weather

Forecasting, Identification & Uncertainty Quantification

Mandar Hemant Chandorkar

Machine Learning in Space Weather

Forecasting, Identification & Uncertainty Quantification

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de Technische Universiteit
Eindhoven, op gezag van de rector magnificus prof.dr.ir. F.P.T. Baaijens,
voor een commissie aangewezen door het College voor Promoties, in het
openbaar te verdedigen op donderdag 14 november 2019 om 16:00 uur

door

Mandar Hemant Chandorkar

geboren te Mumbai, India

Dit proefschrift is goedgekeurd door de promotoren en de samenstelling van de promotiecommissie is als volgt:

Voorzitter		prof.dr.ir. G.M.W. Kroesen	
Promotores	1e Promotor	prof.dr. P.D. Grünwald	Universiteit Leiden
	2e Promotor	prof.dr. U.M. Ebert	
	Copromotor	dr. E. Camporeale	Centrum Wiskunde en Informatica
Leden		prof.dr. M. Owens	University of Reading
		prof.dr. N.J. Lopes Cardozo	
		prof.dr. M. Girolami	The University of Warwick
Adviseurs	1e Adviseur	dr. M. Sebag	Université Paris Sud
	2e Adviseur	dr. C. Furtlehner	INRIA Saclay - Île-de-France

Het onderzoek of ontwerp dat in dit proefschrift wordt beschreven is uitgevoerd in overeenstemming met de TU/e Gedragscode Wetenschapsbeoefening.

The research in this thesis was funded by CWI and received generous financial support for travel from INRIA. It was carried out in the MDG-TAO associate team of the CWI-INRIA International Lab (`https://project.inria.fr/inriacwi/`).

Cover Design: Andong Hu; cover created with using the neural style transfer online tool (`https://tenso.rs/demos/fast-neural-style/`). A NASA image of the Sun-Earth system was styled on the painting *Udnie* by Francis Picabia (1913).

Printing: Ipskamp Printing B.V.

A catalogue record is available from the Eindhoven University of Technology Library.

ISBN: 978-90-386-4903-0

Summary

Machine Learning in Space Weather

Forecasting, Identification & Uncertainty Quantification

The study of variations in the space environment between the Sun and the Earth constitutes the core of space weather research. Plasma ejected by the Sun couples with the Earth's magnetic field in complex ways that determine the state of the Earth's magnetosphere. Adverse effects from space weather can impact communication networks, power grids and logistics infrastructure, all crucial pillars of a civilization that is reliant on technology.

It is important to use data sources, scientific knowledge and statistical techniques to create space weather forecasting and monitoring systems of the future. This thesis aims to be a step towards that goal. The work is organised into the following chapters.

In chapters 4 and 5, we develop probabilistic forecasting models for predicting geo-magnetic time series. Combining ground based and satellite measurements, we propose a gaussian process model for forecasting of the Dst time series one hour ahead. We augment this model with a long short-term memory (LSTM) network and produce six-hour-ahead probabilistic forecasts for Dst.

Quantifying uncertainties in the dynamics of the Earth's radiation belt is an important step for producing ensembles of high fidelity simulations of the magnetosphere. In chapter 6, we infer uncertainties in magnetospheric parameters, using data from probes orbiting in the radiation belts, by combining simplified physical models of the radiation belt with Markov Chain Monte Carlo techniques.

In time-varying systems, it is often the case that cause and effect don't occur at the same time. A prominent example of this time-lagged behaviour is the Sun-Earth system. Particles ejected from the Sun, also called the solar wind, reach the Earth's magnetosphere after a time delay which is dynamic and uncertain. In chapter 7, we propose a novel neural network based method, called Dynamic Time Lag Regression (DTLR), for predicting time-lagged effects of events. We apply the DTLR methodology to the problem of near-Earth solar wind forecasting from heliospheric data.

Acknowledgements

Thank you Enrico for giving me this opportunity, for your guidance, and your belief in me. Michélè and Ute, you have both taught me, in your own ways, what real leadership is about. Cyril, thank you for teaching me that patient rumination and deep thinking are still relevant in today's world. Bala, your wisdom and support elevated the quality of our research, you are the kind of scientist the community needs and I wish you the very best.

I would like to thank Peter Grünwald for his guidance and thoughtfulness during the manuscript preparation process. Peter, you truly are a beacon for young scientists.

Behnaaz, we wont be able to continue our afternoon gossip sessions, but you should keep at it because gossip is known to lead to a long and happy life. Ashutosh, Casper, and Gabriel, your company meant that there was never a dull day at the office. Algo, always remember that the world is a non-linear, non-stationary, chaotic dynamic system at the edge of criticality, so if you have a sample of 100 clean data points and a sample of 100 corrupted data points, you should add noise to the clean data set, in order to equalise the disaster.

Rakesh and Carl(os), I will always remember and miss our hilarious chats, very nice! Hani, never change! Giancarlo, salud to Parisian evenings with Armagnac. Monsieur Philippe, continuez à faire ces sublimes galettes. Nada and Bikkie, you make CWI a great place to work. I would not have gotten through this PhD so smoothly if it wasn't for you guys!

Divya, without you, none of this would have ever happened! Aai and Baba, I am what I am today because of the childhood that you provided me. An upbringing that was full of curiosity, support, travel, wisdom, and love.

Contents

List of Figures

List of Tables

Chapter 1

Outline

The study of variations in the space environment between the Sun and the Earth constitutes the core of *space weather* research. Ionised plasma ejected by the Sun couples with the Earth's magnetic field in complex processes that determine the state of the Earth's magnetosphere. Adverse effects from space weather can impact communication networks, power grids, and logistics infrastructure, which are all crucial pillars of a civilization that is reliant on technology.

It is important to leverage data sources, scientific knowledge and machine learning techniques to create space weather forecasting and monitoring systems of the future. This thesis aims to be a step towards that goal.

1.1 Research Questions

The research questions addressed in this dissertation center around three key themes.

1. *Standard Forecasting*: How can we create probabilistic forecasting systems for key geomagnetic quantities? What is the time horizon for such forecasts? Is it possible to get accurate predictions while increasing the forecast horizons?

2. *Parameter Inference & Uncertainty Quantification*: Are machine learning models and physics models two separate pieces or can they be a

married together in a way which enables identification of unobserved physical parameters?

3. *Dynamic Time Lag Regression*: How can we predict near Earth solar wind speed and its propagation time from solar data? More generally, is it possible to infer using time series data *when* the impacts of events will be felt at their downstream destinations? This is different from standard forecasting which is always concerned with predictions made with a *fixed* time horizon.

1.2 Chapter Outline

The dissertation is organised as follows.

- Chapter 2 provides context for the space weather project and its links to terrestrial weather via historical anecdotes. It highlights impacts of space weather events and speculates why space weather research will become even more important in the coming decades.

- Chapter 3 gives a short introduction to the concepts needed to understand the problems considered in this thesis. Section 3.3 talks about the magnetosphere and the motions of charged particles trapped within it. This is the starting point for the material in chapters 4 to 6. Section 3.2 gives a quick overview about the structure of the Sun, its magnetic field, and the solar wind which is the target application of the method proposed in chapter 7.

- Chapter 4 applies *Gaussian process* (GP) models for making probabilistic forecasts of Earth based geomagnetic quantities and gives a practical methodology for building and evaluating such models. Chapter 5 extends the time horizon of the forecasting models proposed in chapter 4 by proposing a hybrid model based on *long short-term memory* (LSTM) networks and Gaussian processes.

- Chapter 6 proposes a model based on the *least squares support vector machine* (LSSVM) for estimating magnetospheric plasma density. The parameters of the model are estimated by minimizing error with respect to observed data and the physical dynamics of plasma diffusion. Combining this with *Markov chain Monte Carlo* methods, we obtain Bayesian estimates on unobserved parameters of the plasma diffusion system.

- Chapter 7 introduces *dynamic time lag regression* (DTLR) a novel supervised regression framework which captures probabilistic and dynamic propagation time delays between time series. We propose a solution methodology for the DTLR setting and provide a theoretical framework for understanding its convergence. The DTLR methodology is applied to the forecasting of solar wind speed from solar magnetic data.
- Chapter 8 discusses the progress made in the thesis and offers perspectives for future research.

Chapters 4 and 5 are based on published journal articles while chapters 6 and 7 are based on work that is in peer review or in preparation for publication.

Chapter 2

Historical Perspectives

Weather forecast for tonight: dark.

George Carlin

Earth, *Wind*, *Fire* & *Water*, the *classical elements* were the basis for understanding our environment during antiquity. Modern science, based on experiments, has taken a very different view of the world, one based on atoms, fundamental particles, and states of matter. However, we could argue that the classical elements were a more philosophical idea that distilled our everyday experiences with nature. In fact, many ancient cultures such as Hellenistic Greece, Babylonia, Japan, Tibet, China, and India had similar lists of four or five elements. These civilizations had very different views on the properties of these elements and how they related to natural phenomena, quite often these links were mythological. The obvious way in which people experienced the classical elements was through weather systems.

From the seasons to daily weather variations, nature's elements drive and shape our lives. Sometimes weather has had a direct impact on entire populations. One example was the failed Mongol invasions of Japan in 1274 and 1281. In both attacks, the Mongol fleets were almost entirely destroyed by storms called *kamikaze* (translates to divine wind). Although some attacking Mongol forces did manage to land during the 1274 campaign and outnumbered the defending armies, they were still defeated by Samurai clans with superior knowledge of the terrain.

Figure 2.1: The Mongol fleet destroyed in a typhoon, 1847. *Source*: Kikuchi Yōsai / Tokyo National Museum (Public domain)

The invading fleet of 1281 was composed of 'more than four thousand ships bearing nearly 140, 000 men' [McClain, 2002, pg. 17], the scale of which was eclipsed only by the allied invasion of Normandy in 1944. The fleet was a hastily assembled, consisting of ships which were not suitable for the harsh waters between Japan and Korea. The Japanese had built two metre high walls in the intervening period and the invading fleets were forced to stay in sea for months. After their supplies were diminished, powerful kamikaze winds destroyed them entirely (an artist's view of the event is illustrated in figure 2.1). The failed invasions were a blow to the idea of Mongol supremacy in Asia and the Mongols never attempted an invasion of Japan since.

We now know that weather phenomena are caused by a combination of air pressure, temperature, and moisture differences between one place and another. The angle of the Sun's rays changes with latitude, and these variations create very different temperature trends from the poles to the equator. These differences in temperature lead to large scale air currents which create complex weather systems and climate patterns which we see across the world. But weather phenomena are hardly exclusive to planet Earth.

The Final Frontier

Weather phenomena occurring on other planets have been observed even before the beginning of the space age. Jupiter's *great red spot*, a huge storm, has been continuously observed since 1830 [see Britannica].

Saturn's *great white spot* is a recurring storm system which was first used by Asaph Hall to determine the period of the planet's rotation [Wikisource, 2014]. In the 20^{th} century, missions such as the Hubble space tele-

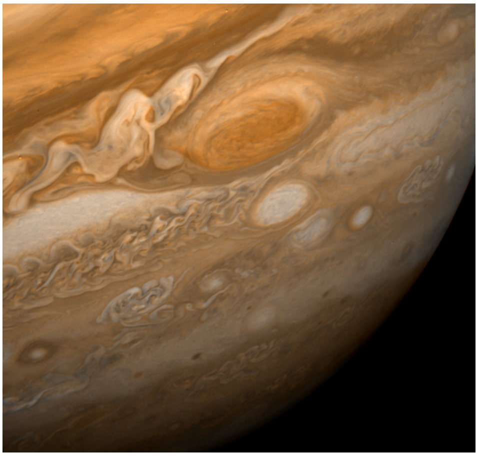


Figure 2.2: Jupiter's Great Red Spot in February 1979, photographed by the unmanned Voyager 1 NASA space probe. *Source*: NASA (Public domain)

scope, Voyager, Cassini, and others have shown storms and other weather phenomena on planetary bodies like Venus, Mars, Neptune, and Titan.

The principles behind many planetary weather phenomena are very similar. Each planet has a different atmosphere so weather phenomena in the solar system can have very different characteristics. Extra-terrestrial weather is just as complex and mind boggling as the weather we observe on Earth, and its scale is certainly much larger than we are used to.

However, planetary weather is just one side of the puzzle. Venturing into our cosmic neighbourhood, our solar system has another kind of weather system that has begun to be probed only very recently.

A Gust of Wind from the Heavens

During the last week of August 1859, several spots appeared on the surface of the Sun. Southern auroral displays were observed on August 29 as far north as Queensland, Australia. Just before noon on September 1, British astronomer Richard Carrington observed a 'white light flare' from a group of sun spots. He created a sketch of his observations which is seen in figure 2.3. Carrington's observations were independently verified by British publisher and astronomer Richard Hodgson; both of them sent their reports to the *Monthly Notices of the Royal Astronomical Society.*

September 1-2 1859 saw some remarkable events occur around the world. Auroral displays were observed all around the world, even in low latitude places such as Colombia [Cárdenas et al., 2016]. Auroras above the rocky mountains in the U.S were so bright that they woke up gold miners who began preparing breakfast thinking it was morning [Sten F. Odenwald]. In the northeastern U.S, people could read the newspaper by the aurora's light [Lovett].

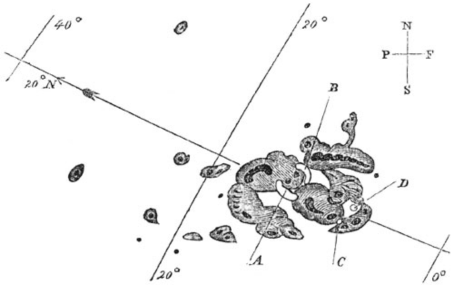


Figure 2.3: Sunspots of September 1, 1859, as sketched by Richard Carrington. A and B mark the initial positions of an intensely bright event, which moved over the course of five minutes to C and D before disappearing. *Source*: Richard Carrington (Public domain)

The telegraph network in Europe and North America failed. Some operators experienced electric shocks [Board, 2008, pg. 13] while in some cases even telegraph equipment that was disconnected from the power supply could be used to transmit messages [Carlowicz and Lopez, 2002, pg. 58].

Based on global reports and observations taken by Scottish physicist Balfour Stewart at the Kew observatory in London, Carrington was able to connect events observed on Earth to what he saw on the Sun on the 1st of September [Clark and Clark, 2007]. His assertion was corroborated by other observers in the scientific community.

The storm of 1859, later known as the *Carrington event*, was in some ways the genesis of the *Space Weather* domain; however the actual term was coined much later in the 1950s. Although scientists had observed sunspots and their links to magnetic field variations on the Earth earlier, the *Carrington event* was a concrete example of how activity on the Sun could have potentially dramatic effects on the Earth.

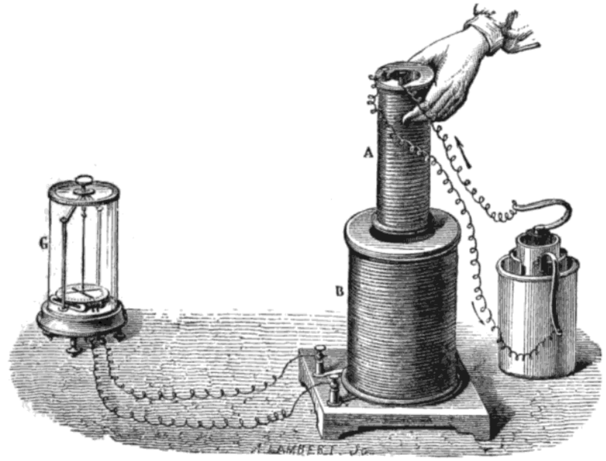


Figure 2.4: One of Faraday's 1831 experiments demonstrating induction. The liquid battery (right) sends an electric current through the small coil (A). When it is moved in or out of the large coil (B), its magnetic field induces a momentary voltage in the coil, which is detected by the galvanometer (G). *Source*: J. Lambert (Public domain)

Space Weather

How do spots and ejections from the Sun produce bright lights and currents on Earth? During Carrington's time, progress in the fledgling science of electromagnetism had picked up and enabled some understanding of the link between solar outbursts and geomagnetic phenomena. Faraday's induction experiment in 1831 (figure 2.4) demonstrated that varying magnetic fields could induce electrical currents in copper wires.

It took approximately a century from the *Carrington event* for a theoretical understanding of Space Weather phenomena to develop. Maxwell's equations of electromagnetism [Maxwell, 1865] published in 1864 gave scientists the mathematical tools to model the motions of charged particles in electric and magnetic fields, and the variations in the fields themselves due to their motions.

Rapid progress was made in the 20^{th} century in modelling the motions of charged particles trapped in the Earth's magnetic field, in the area of *plasma physics.* Plasma was the name given to the state of matter which existed as an ionised gas. From the scientific advances made in space weather and plasma physics, we know that the Carrington white light flare in 1859 was accompanied by a large release of energetic plasma from the Sun's atmosphere, also known as a *coronal mass ejection* (CME). The CME associated with the Carrington event was particularly energetic and compressed the Earth's magnetic field causing currents to flow in conducting materials like telegraph equipment.

Space weather started gaining relevance with the rise of space missions, and increasing reliance on communications networks. The risks posed by space weather to space faring assets like satellites meant that understanding and forecasting space weather events became especially important; although there is still much progress yet to be made.

Impacts

The Quebec power grid failure of 1989 [Kappenman et al., 1997] during a geomagnetic storm event showed that intense space weather events like the one observed in 1859 could cause significant damage to communications, energy, and technological infrastructure that is so crucial to the working of modern civilization.

It is now widely accepted that space weather events can adversely impact satellite and communication infrastructure, airline industry, navigation systems and the electric power grid [Board et al., 2009, Cannon et al., 2013, Bothmer and Daglis, 2007, Baker et al., 2004]. To protect our technological systems and humans in deep space exploration, it is necessary to have an advance knowledge of the changes in space weather that can pose potential threats.

The Future

The solar storms observed in the 20th and 19th centuries are only one part of the picture. It is now increasingly likely that private companies will be making significant inroads into space travel for business goals. Companies such as SpaceX and BlueOrigin aim to make space travel cheaper and more accessible so that human beings can live and work in space or other planets in the solar system, potentially starting a second space age.

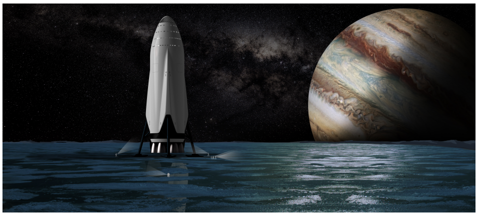


Figure 2.5: Artist's impression of the SpaceX Interplanetary Starship on the Jupiter's moon Europa *Source*: Space Exploration Technologies Corp. [CC0]

This drastic move to become a multi-planetary species will bring with it risks to human life and equipment. These risks come in the form of severe magnetic storms, solar flares, and ejections of charged particles, which must be anticipated if we want to become a successful space faring race.

In order to design resilient technological systems for the new space age, we need to make progress in understanding and anticipating space weather phenomena. Physical theories about space plasmas needed to be combined with the data collected from space missions. The rapid rise of hardware, software, and data storage; the deluge of space mission data, and the advent of machine learning techniques means that we are in a unique position to take strides towards our space goals.

Chapter 3

Background

Space weather is the branch of physics that studies the time varying phenomena in the solar system. The principal driver of space weather phenomena is the Sun, specifically its magnetic field variations and the *solar wind*. The effect of solar variations on the planetary environment are caused by the coupling between solar wind particles and the magnetic field produced by the Earth. This chapter gives a semi-quantitative treatment of various scientific ideas relevant to space weather research.

Table 3.1 provides the reader with a condensed guide to this dissertation. It connects the content presented in this chapter with the main research problems analyzed in the later chapters and provides recommended prerequisite reading for each chapter.

Chapters	**Themes**	**Recommended Reading**
4 & 5	Forecasting of geomagnetic index Dst.	§ 3.3.2, § 3.3*, § 3.1*
6	Inference of radiation belt parameters.	§ 3.3.1, § 3.1* § 3.3*
7	Forecasting of near Earth solar wind speed using solar data.	§ 3.2.2, § 3.2.3, § 3.2*

Table 3.1: Dissertation Guide. Asterisk* denotes optional material.

Section 3.1 describes space plasmas and their properties. Section 3.2 provides some background about the Sun and the solar wind which is the driver for all space weather phenomena. This is used in the solar wind prediction task considered in chapter 7. Section 3.3.1 introduces the plasma diffusion model (equation (3.10)) and its simplified radial diffusion system (equation (3.14)) which is used as the underlying physical model for chapter 6. Section 3.3 introduces the *magnetosphere*, giving context for chapters 4 and 5.

3.1 Space Plasma

Plasma, also known as the fourth state of matter due to its properties that differentiate it from the conventional gaseous state, is ubiquitous throughout the visible Universe. Plasma is a gas which is composed of roughly equal number of positive and negatively charged particles, a property known as charge *quasi-neutrality*. The term quasi-neutral is used because although the gas has almost equal amounts of positive and negative charges, the mixture is electromagnetically active. Due to incomplete charge shielding, long range electromagnetic fields play a big role in the dynamics of plasma.

Debye Length

In a quasi-neutral plasma, due to the presence of partial electric shielding the potential due to the charges now takes the well known Debye form

$$\phi(r) = \frac{q}{4\pi\epsilon_0 r} e^{-\frac{r}{\lambda_d}} \ , \tag{3.1}$$

where r is the spatial distance with respect to the charge and ϵ_0 is the *permittivity* of vacuum. The electric potential decays with the Debye length scale λ_d at which a balance between thermal vibrations which can disturb quasi-neutrality, and electrostatic forces due to charge separation, is achieved. The Debye length scale depends on the electron temperature and plasma density.

$$\lambda_d = \sqrt{\frac{\epsilon_0 k_b T_e}{n_e e^2}} \tag{3.2}$$

In equation (3.2) above, the Debye length scale is expressed in terms of the *Boltzmann constant* k_b, the electron temperature T_e, free space permittivity

ϵ_0, and electron charge e. One can visualise the positively charged ions having a cloud of electrons shielding them at the distance of λ_d.

It is also possible to take into account the shielding effect of the ions. The effective Debye length is now expressed as an addition of two terms: one for electrons (equation (3.2)) and a similar term for the ions by replacing T_e for the ion temperature T_i ($n_i \approx n_e$).

Plasma Parameter

Consider a Debye sphere of radius λ_d. This sphere contains $N_e = \frac{4}{3}\pi\lambda_d^3 n_e$ electrons. The plasma parameter g is defined as N_e^{-1}. Rewriting this, we can say:

$$g \sim \sqrt{\frac{n_e}{T_e}} \ .$$

The description of plasma used in many applications in space is applicable when $g \ll 1$. In this situation the Debye shielding is significant, and the quasi-neutral plasma obeys collective statistical behavior. The plasma parameter g also correlates with the collision frequency. The collisions in plasma increase with increasing density and decreasing temperature, and if $g \longrightarrow 0$ the plasma becomes nearly collisionless. The collisionless property helps in making simplifying assumptions about plasma dynamics and serves as the starting point for the *adiabatic* theory of plasma motions in the Earth's magnetosphere which will be discussed in section 3.3.1.

3.2 Sun & the Solar Wind

The Sun is an almost perfectly spherical ball of plasma which is the the center of our solar system and the only source of light and energy for all living and meteorological processes on Earth. Apart from terrestrial weather, the Sun is also the primary driver of space weather which results from the interaction between the solar wind and planetary magnetospheres.

3.2.1 Structure

Figure 3.1 shows a cross section of the Sun with various layers. We give a brief description of them below.

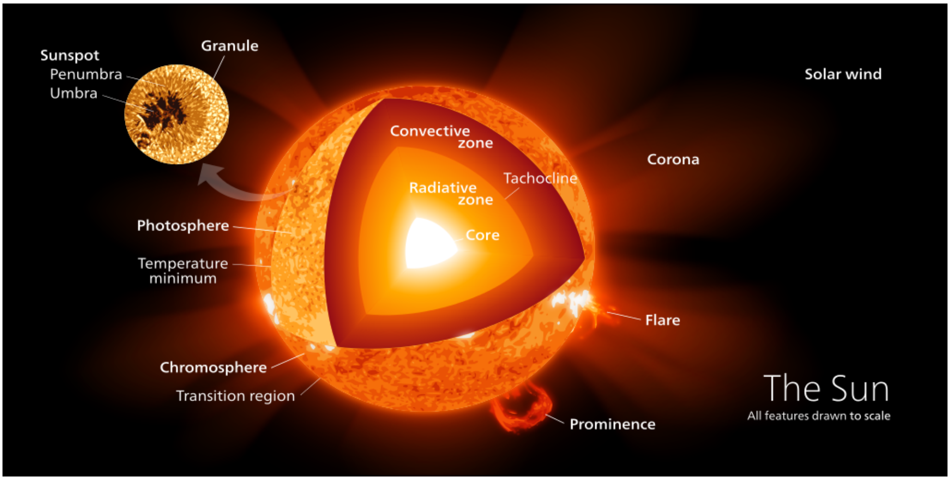


Figure 3.1: Cross section of the Sun
Source: Kelvinsong [CC BY-SA 3.0 (`https://creativecommons.org/licenses/by-sa/3.0`)]

Core: The core of the Sun is the site for the thermonuclear fusion reactions which produce its energy. It extends from the center to about $20 - 25\%$ of the solar radius [García et al., 2007]. It has a temperature close to 1.57×10^7 K and a density of $150\,\mathrm{g/cm^3}$ [Basu et al., 2009]. Nuclear fusion in the core takes place via the well known *proton-proton chain* (pp).

Radiative Zone: The radiative zone extends from 25% to 70% of the solar radius. The nuclear reactions in the core are highly sensitive to temperature and pressure. In fact, they are almost shut off at the edge of the core. In the radiative zone, energy transfer takes place via photons (radiation) which bounce around nuclei until they reach the convective zone.

Convective Zone: The convective zone lies between 70% of the solar radius to a point close to the solar surface. Density decreases dramatically going from the core to the radiative zone and subsequently the convective zone. In this region, the solar material behaves more like a fluid. Due to the temperature gradient which exists across it, the primary source of transport is here via convection.

Photosphere: The photosphere is the visible 'surface' of the Sun, since the layers below it are all opaque to visible light. A layer of about 100 km thickness, the photosphere is also the region from where sunlight can freely escape into space. The photospheric surface has a number of features i.e. sunspots, granules and faculae. Sunspots (see section 3.2.3) are magnetic regions where the solar material has a lower temperature compared to its surroundings. Magnetic field lines are concentrated in sunspot regions, and the field strength in sunspots can often be thousands of times stronger than the on the Earth.

Chromosphere: The Chromosphere extends for a distance of almost 5000 km after the photosphere. The chromosphere is known for the existence of features called spicules and prominences. The chromosphere has a red colour which is generally not visible due to the intense light given off by the photosphere but can be observed through a filter centered on the Hydrogen $H\alpha$ spectral line.

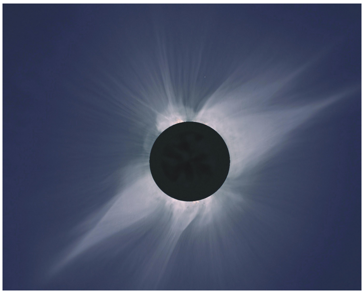


Figure 3.2: Sun's corona captured during a solar eclipse. *Source*: Steve Albers, Boulder, CO; Dennis DiCicco, Sky and Telescope; Gary Emerson, E. E. Barnard Observatory

Solar Transition Region: A thin (100 km) region between the chromosphere and the solar corona where the temperature rises from about 8000 K to 5×10^5 K, the solar transition region might not be well defined at all altitudes; however its existence is evidenced by a bifurcation of the dynamics of the solar plasma. Below the transition region, the dynamics is dictated by gas pressure, fluid dynamics, and gravitation while above the region, the dynamics is dictated more by magnetic forces.

Corona: An aura of plasma around the Sun that extends millions of kilometers into space, the corona can be observed during a total solar eclipse (figure 3.2) or with a coronagraph. The temperature of the corona is dramatically higher than the photosphere and chromosphere. The average temperature can range from 1×10^6 K to 2×10^6 K while in the hottest regions it can be as high as 2×10^7 K [Erdélyi and Ballai, 2007]. Although the reason for this dramatic increase is still not well understood, there exist various explanations using concepts of magnetic reconnection [Russell, 2001, Erdélyi and Ballai, 2007] and Alfvén waves [Alfvén, 1947]. There is a critical height in the corona, known as the *source surface*, below which the magnetic field controls the plasma completely. Above it the plasma carries the magnetic field with it into the interplanetary medium.

3.2.2 Solar Wind & Heliospheric Magnetic Field

The idea that the Sun was ejecting charged particles outwards into space was first hinted at after the solar storm of 1859 by Richard Carrington [Cliver and Dietrich, 2013] and later by George FitzGerald [Meyer-Vernet, 2007]. Arthur Eddington, in a footnote of an article about the comet Morehouse in 1910, was the first to suggest the existence of the solar wind, without naming it so [Durham, 2006].

In the 1950s, studies of the anti-solar orientation of the ion tails of Halley's comet led to the theory of solar corpuscular emission [Biermann, 1952, 1957, 1951]. Parker [1958b, 1960, 1965] argued that the corona cannot remain in hydrostatic equilibrium and that supersonic expansion of the corona is responsible for the outward expulsion of charged particles, which the author referred to as the *solar wind*.

Parker [1958b] also proposed a spiral model for the *Heliospheric Magnetic Field* (HMF) and suggested that the solar wind carried with it the solar magnetic field. The Parker model was further supported its ability to explain the effect of the HMF on the modulation of galactic cosmic rays and their measured intensities close to the Earth [Parker, 1958a]. In 1959 the Soviet spacecraft Luna 1 was the first to directly observe the solar wind and measure its strength [Harvey, 2007]. Subsequently, the Mariner 2 mission recorded properties of the positive ion component of the solar wind and confirmed the Parker spiral HMF model [Neugebauer and Snyder, 1966].

The structure of the HMF is central to explaining the formation and propagation of the solar wind. The HMF in steady state points radially outward and rotates with the Sun, producing an *Archimedean spiral* structure as postulated in Parker [1958b] and shown schematically in figure 3.3. Photospheric observations of the magnetic field (see Global Oscillation Network Group `https://gong.nso.edu`) are often extrapolated to compute approximations to the coronal HMF topology. There exist a number of techniques used to perform such extrapolations: potential field based methods such as *Potential-Field Source Surface* (PFSS) [Schatten et al., 1969, Altschuler and Newkirk, 1969], PFSS variants such as *Potential-Field Current Sheet* (PFCS) [Schatten, 1971], *Current-Sheet Source Surface* (CSSS) [Zhao and Hoeksema, 1995], and several others. Apart from potential based models, there exist more involved techniques based on Magnetohydrodynamics (MHD) such as *Magnetohydrodynamics Around a Sphere* (MAS) [Linker et al., 1999], ENLIL [Odstrčil et al., 1996, Odstrčil and Pizzo, 1999a,b, Odstrčil, 2003, Odstrčil et al., 2004] and EUHFORIA [Pomoell and Poedts, 2018].

The HMF can be seen as a combination of two components: the poloidal magnetic field and the toroidal magnetic field. The two fields often exchange energy between themselves over the course of several years in a cyclical phenomenon known as the *solar cycle* (section 3.2.3). Interested readers can read Owens and Forsyth [2013] for an in-depth review on the phenomena that drive the HMF.

The expansion of the coronal magnetic field leads to an eventual opening of field lines at the source surface (see figure 3.3) and the ejection of the solar wind. This hot plasma consists mostly of protons, electrons and a small number of helium and heavy ions. The solar wind spirals outwards in all directions, carrying with it the magnetic field. Close to the Earth's magnetosphere, this wind has a nominal speed of about $400\,\mathrm{km\,s^{-1}}$ while its high speed component has an average velocity of $\sim 700\,\mathrm{km\,s^{-1}}$ (figure 3.4).

Near Earth Measurements

The solar wind has the heliospheric magnetic field *frozen in*[1], and as it propagates in the interplanetary medium, it carries the solar magnetic field with it [Alfvén, 1942, 1943]. Important solar wind quantities such as: 1. solar wind speed, 2. proton density, and 3. magnetic field strength are recorded at the well known L_1 *Lagrangian point* where the gravitational fields of the Earth and the Sun approximately balance out.

3.2.3 Sunspots & Solar Cycle

Sunspots are temporarily occurring regions on the Sun's photosphere that appear as dark spots. They are areas of magnetic field concentration

---

[1] the flux of the magnetic field going through a surface that moves with the solar wind (in a Lagrangian manner) is constant

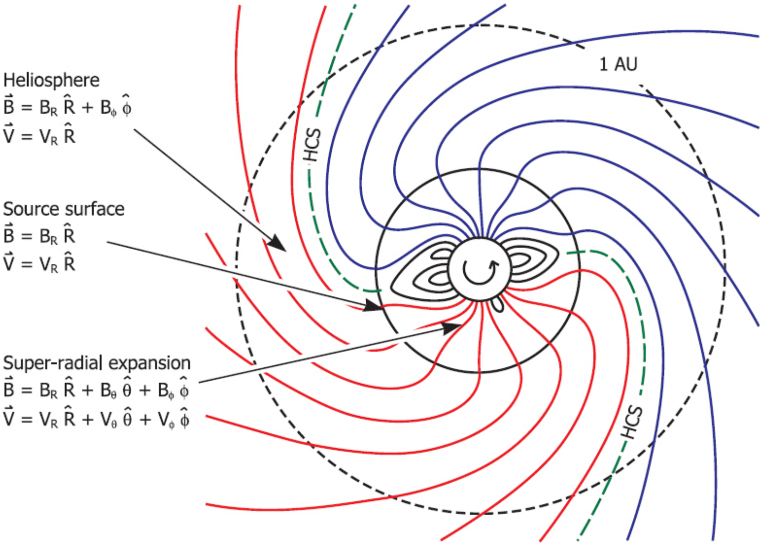


Figure 3.3: An illustration of the Heliospheric Magnetic Field in the *ecliptic plane*. In the heliosphere, rotation of the HMF foot points within a radial solar wind flow generates an azimuthal component of the HMF, B_ϕ, leading to a spiral geometry. Red and blue lines, showing regions of opposite polarity, are separated by the heliospheric current sheet (HCS), shown as the green dashed line. Reprinted by permission from Springer Nature Customer Service Centre GmbH: Springer Nature, Living Reviews in Solar Physics [Owens and Forsyth, 2013], © 2013

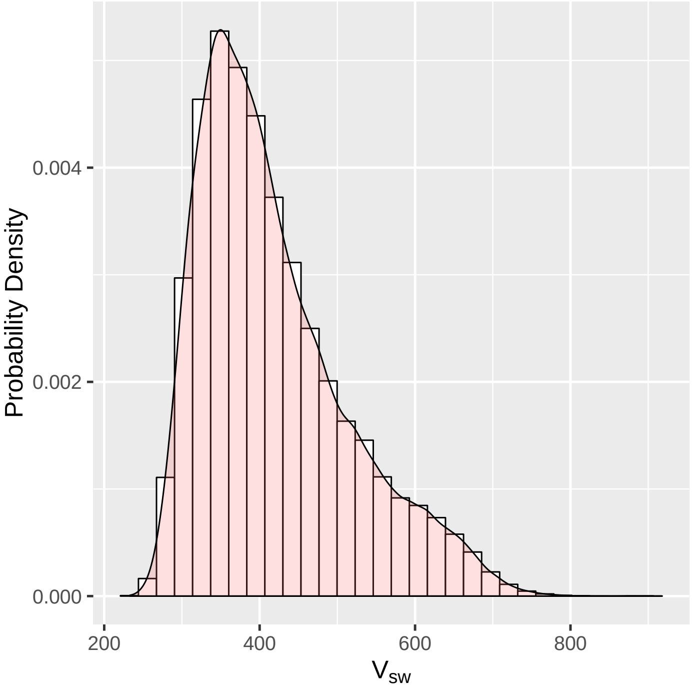


Figure 3.4: Distribution of solar wind speed recorded at 1 au for the time period 2008 – 2018, *Source*: OMNI data set (`https://omniweb.gsfc.nasa.gov/ow.html`)

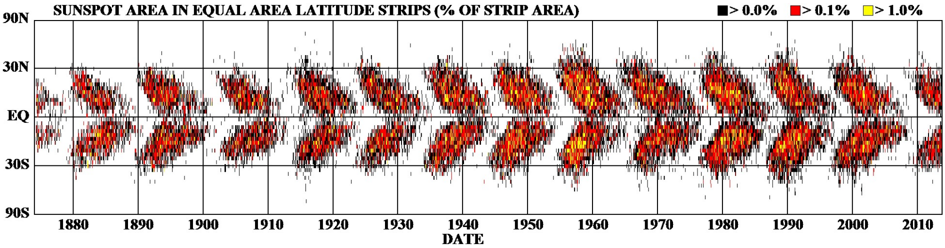


Figure 3.5: The sunspot butterfly diagram. *Source*: By Con-struct - Royal Observatory, Greenwich, data prepared on: `http://solarscience.msfc.nasa.gov/greenwch`, [CC BY-SA 3.0 (`https://creativecommons.org/licenses/by-sa/3.0`)]

where the field lines often ‘puncture’ the solar surface inhibiting convection and producing regions with lower temperatures than the surroundings. Sunspots generally last anywhere between a few days to a few months. They can occur in pairs or groups and can accompany other phenomena such as *coronal loops*, *prominences*, and reconnection events.

Since the 19^{th} century the number of sunspots on the Sun’s surface have been recorded as the *sunspot number* (SSN). Sunspots populations increase and decrease, thereby behaving as markers for solar activity levels. The cyclical behavior of sunspot populations is called the *sunspot cycle* or *solar cycle* (figure 3.5).

Figure 3.5 depicts how the area occupied by sunspots changes with solar latitude and time. During the start of a solar cycle (solar minimum), sunspots start appearing at higher latitudes. Over the course of the cycle, they move towards the equatorial regions and their number increases to some maximum (solar maximum). Towards the end, the number of sunspots diminishes and the entire cycle starts over. This repetitive behavior happens over approximately 11 years.

Because sunspots are magnetic phenomena, the solar cycle represents cyclical behavior of the HMF. During solar minimum, the poloidal component of the solar magnetic field is at its strongest and it is the closest it can get to a magnetic dipole configuration. Towards solar maximum, energy is transferred from the poloidal component to the toroidal component, resulting in complex field configurations which are evidenced by larger numbers of sunspot clusters.

The solar cycle also gives rise to variations in solar irradiance [Willson et al., 1981]. Between 1645 and 1715, very few sunspots were observed, a period known as the *Maunder minimum.* This coincided with lower than average temperatures in Europe, which was called the *little ice age.* Although the Maunder minimum was a period of lower solar irradiance, recent research [Owens et al., 2017] has demonstrated that this was neither the only factor nor the most significant in causing lower than average temperatures during the little ice age.

In chapter 7, the *sunspot number* data as well as the *flux tube expansion* factor (f_S or FTE) and the magnetic field strength computed by the CSSS model will be used to create a input data set for building the *dynamic time lag regression* model proposed therein. Using the input parameters, the DTLR model provides an estimate for the near Earth solar wind speed as well as the propagation time. Measurements of the solar wind speed will also be used in chapters 4 and 5 as inputs to the Dst forecasting models applied therein.

3.3 Magnetosphere

The Earth's magnetosphere (figure 3.6) is a region surrounding the planet where its magnetic field dominates the *interplanetary magnetic field.* The Earth's magnetic field shields the atmosphere and terrestrial life from the impact of the solar wind.

As the solar wind approaches the Earth, it is slowed down and deflected by the Earth's magnetic field. Since the solar wind is supersonic when it arrives and slows down to subsonic levels, a shock wave is generated in the process (*bow shock*). Much of the solar wind kinetic energy is converted to thermal energy when it crosses the *bow shock* into the *magnetosheath.* The *magnetosheath* spans from the *bow shock* to the *magnetopause.* The *magnetopause* is the outer boundary of the Earth's magnetic shield. Its location is $\sim 10 R_E$ ($R_E = 6372\,\mathrm{km}$, the radius of the Earth).

Earth's magnetic shielding is not perfect, and some particles manage to get trapped inside the cavity of the *magnetosphere.* This region of trapped plasma is known as the the *van Allen radiation belts.* Particles trapped in the radiation belts execute complex motions which can be approximately

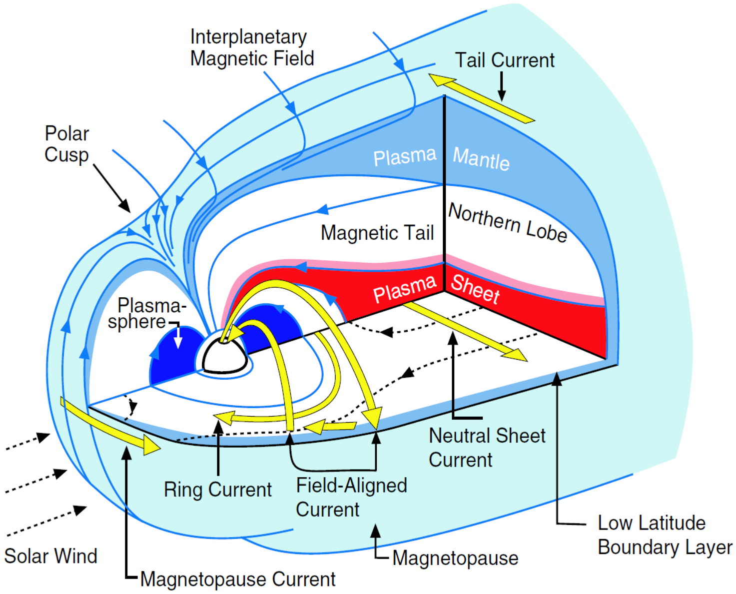


Figure 3.6: Three-dimensional cutaway view of the magnetosphere. The light blue outer surface is the magnetopause, and its boundary layers are shown in darker blue. Magnetic field lines are shown in blue, electric currents in yellow. The polar region where the magnetic field lines converge is the polar cusp. The bow shock has been omitted for clarity. Reprinted by permission from Springer Nature Customer Service Centre GmbH: Springer Nature, Space Science Reviews [De Keyser et al., 2005], © 2005.

modelled using ideas from adiabatic theory and diffusion described in section 3.3.1 below. The *plasmasphere* is the inner region of the radiation belts which contains cold, dense plasma. The portion of the magnetosphere facing away from the Sun (called the *nightside*) is stretched out in a tail-like shape by the deflected solar wind, hence referred to as the *magnetotail*. The *magnetotail* has an approximate extent of up to $1000 R_E$.

3.3.1 Particle Motions & Adiabatic Theory

This section gives a quick introduction to the theory of charged particle motions in the magnetosphere. The reader may refer to Roederer [1970] for an in depth treatment of this subject. To understand the motions of charged particles in the *magnetosphere*, the role of electric and magnetic forces must be understood.

It is well known from classical electromagnetism that the force exerted on a particle with charge q by a magnetic field $\mathbf{B}$ and an electric field $\mathbf{E}$ is given by the well known *Lorentz force* (equation (3.3)).

$$\mathbf{F} = m\frac{d\mathbf{v}}{dt} = q\mathbf{E} + q\mathbf{v} \times \mathbf{B} \tag{3.3}$$

The first component of equation (3.3) ($q\mathbf{E}$) is either parallel or opposite to the local electric field depending on the charge of the particle. The second component $q\mathbf{v} \times \mathbf{B}$ involves a vector cross product so it is always perpendicular to the plane spanned by vectors $\mathbf{v}$ and $\mathbf{B}$. In order to understand its effects, we can decompose the particle velocity in two components; $v_{||}$ parallel to $\mathbf{B}$ and $v_{\perp}$ perpendicular to $\mathbf{B}$. If $\mathbf{E} = \mathbf{0}$, then the particle executes a circular motion with properties shown in equation (3.4). Here ρ is the gyroradius and ω is the gyrofrequency or cyclotron frequency. In the case where $v_{||} \neq 0$, the trajectory is helical.

$$\frac{v_{\perp}^2}{\rho} = \frac{qBv_{\perp}}{m} = \omega v_{\perp} \tag{3.4}$$

Apart from the gyro motion, there are some important drift forces that significantly influence particle motions.

- Electric field drift: If $\mathbf{E}$ has a component $E_{\perp}$ perpendicular to $\mathbf{B}$, the electric field accelerates and decelerates the particle in the two hemispheres of the orbit. The orbit becomes a distorted circle, and the particle drifts in a direction perpendicular to the electric field with a velocity $\mathbf{v_d} = \mathbf{E} \times \mathbf{B}/B^2$.

- Magnetic gradient drift: When the magnetic field varies in space (as is the case of the Earth), a gradient in the field strength in the direction perpendicular to $\mathbf{B}$ gives rise to a gradient drift velocity given by $\mathbf{v_g} = \frac{1}{2}mv_{\perp}^2\mathbf{B} \times \frac{\nabla\mathbf{B}}{aB^3}$.

- Magnetic curvature drift: If the magnetic field has a curvature, this creates an additional drift motion with velocity $\mathbf{v_c} = \frac{m v_{\|} \mathbf{B} \times (\hat{\mathbf{b}} \cdot \nabla) \hat{\mathbf{b}}}{q B^2}$ $(\mathbf{b} = \frac{\mathbf{B}}{B})$.

The equations of motion for charged particles in the general case of spatially varying electric and magnetic fields do not admit closed-form solutions. The motions are generally complex and require lengthy numerical integrations to be resolved.

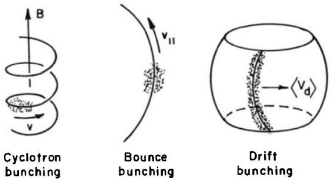


Figure 3.7: The periodic components of the motion of trapped particles. Reprinted/adapted by permission from Springer Nature Customer Service Centre GmbH: Springer Nature, Dynamics of geomagnetically trapped radiation by J. G. Roederer © 1970.

The *guiding center* approximation helps us to decompose particle motions into three periodic components (figure 3.7): 1. gyration around magnetic field lines, 2. bounce motions between magnetic north and south poles, and 3. equatorial drift of electrons and protons, each with its own time scale.

Adiabatic Invariants

When a physical system with periodic motion is varied slowly as compared to the time period of its periodicity, the transformation can be characterized as *adiabatic.* Formally speaking, for systems which are described by Hamiltonian dynamics, we can write the equations of motion in terms of the *canonical position* q, the *canonical momentum* p, external parameters θ, and the system's *Hamiltonian* $\mathcal{H}(q, p|\theta)$:

$$\begin{aligned} \dot{q} &= \frac{\partial \mathcal{H}}{\partial p} \\ \dot{p} &= -\frac{\partial \mathcal{H}}{\partial q} \ . \end{aligned} \tag{3.5}$$

If the system shown in equation (3.5) executes a periodic motion in the q, p phase space, it admits an *adiabatic invariant* A given in equation (3.6).

$$A = \oint p dq \tag{3.6}$$

The quantity A would remain approximately constant if the external parameters θ were varied adiabatically (i.e. if changes in θ happen over a time period much greater than the period of oscillation of the system).

Applying the idea of adiabatic invariance to charged particle motion in the *magnetosphere*, it is possible to associate one adiabatic invariant with each periodic motion i.e. gyromotion $\mathcal{M}$, bounce J, and drift Φ (equation (3.7)).

$$\mathcal{M} = \frac{1}{2}\frac{mv_\perp^2}{|\mathbf{B}|} \tag{3.7}$$

$$J = \oint m_0 v_{||} ds \tag{3.8}$$

$$\Phi = \oint v_{\text{drift}} \cdot r d\phi = \int \mathbf{B} d\mathbf{S} \tag{3.9}$$

The first invariant $\mathcal{M}$ is associated with the Larmor gyration - it is the magnetic moment of the current generated by the circular motion of the particle around the field line.

The second invariant J is associated with the bounce motion between the two magnetic mirrors near the north and south poles (the quantity s is an appropriately chosen arc length coordinate along the bounce trajectory). The bounce motion between the magnetic poles can be explained by the conservation of the particle energy $\frac{1}{2}mv_{||}^2 + \frac{1}{2}mv_\perp^2$ and the first invariant $\mathcal{M}$. Because field strength $|\mathbf{B}|$ increases near the poles, $v_\perp$ also increases to conserve $\mathcal{M}$; however, to conserve energy, $v_{||}$ decreases until the particle can no longer move farther along the field line (and bounces back).

The third invariant Φ, associated with equatorial drift motion, is actually the magnetic flux through the barrel shape envelope of the particle drift. A particle's guiding magnetic field line can be identified by its radial position r and its longitude ϕ. The magnetic flux of the drift can then be computed by integrating over ϕ.

Associated with each adiabatic invariant is a timescale which determines how easily its conservation can be violated. The timescales for $\mathcal{M}$, J, and Φ are the time periods of the gyromotion, bounce motion, and equatorial drift motion respectively. Since it takes much a longer time for the particles to complete a drift motion around the Earth as compared to bounce and gyromotion (in that order), the invariance of Φ is most easily violated - a fact which is used in the simplification of the Fokker-Planck diffusion system described below.

Plasma Diffusion

Because we consider populations of charged particles, it is natural to employ some kind of distribution based picture for magnetospheric plasma. The adiabatic invariants give us a phase space or coordinate system by which we can express quantities of interest.

The main quantity of interest in this case is the *phase space density* $f(t, \mathcal{M}, J, \Phi)$ which is a function of time and three invariants. The *phase space density* tells us the number of particles in a particular region of the phase space, and at a particular point of time.

Diffusion behavior arises when one or more of the invariants are violated, which can happen due to a number of reasons such as: 1. non-adiabatic variations of the magnetic field, 2. external forces, 3. interaction with electromagnetic waves, and 4. collisions with atmosphere/ionosphere. The plasma diffusion system [Schulz and Lanzerotti, 1974] can be written as a generalized Fokker-Planck system as shown in equation (3.10).

$$\frac{\partial f}{\partial t} = \sum_{p,q=1}^{3} \frac{\partial}{\partial J_p} \left(\kappa_{pq} \frac{\partial f}{\partial J_q} \right) \tag{3.10}$$

$$J_1 = \mathcal{M} \tag{3.11}$$

$$J_2 = J \tag{3.12}$$

$$J_3 = \Phi \tag{3.13}$$

It is possible to simplify this system by considering the two main categories of diffusion: radial diffusion and pitch angle diffusion. Radial diffusion allows particles to move farther or closer to the Earth, and pitch angle[2]

---

[2] pitch angle being the angle between particle velocity $\mathbf{v}$ and the magnetic field $\mathbf{B}$

diffusion moves the magnetic mirror points along the field lines.

Rewriting $\Phi \propto \frac{1}{\ell}$, the third invariant can be expressed in terms of the *drift shell* l (larger value of l implies greater distance from the Earth). The radial diffusion system can be obtained from equation (3.10) by keeping $\mathcal{M}$ and J fixed, considering diffusion in ℓ (violation of Φ invariance), and by approximating pitch angle diffusion as a loss process [Walt, 1970, Roederer, 1970]. The resulting system is shown in equation (3.14).

$$\frac{\partial f}{\partial t} = \ell^2 \frac{\partial}{\partial \ell} \left(\frac{\kappa(\ell, t)}{\ell^2} \frac{\partial f}{\partial \ell} \right)_{\mathcal{M},J} - \lambda(\ell, t) f \tag{3.14}$$

The first term on the right hand side, $\ell^2 \frac{\partial}{\partial \ell} \left(\frac{\kappa(\ell,t)}{\ell^2} \frac{\partial f}{\partial \ell} \right)_{\mathcal{M},J}$, models diffusive phenomena in Φ but is expressed in the drift shell coordinate ℓ. Pitch angle diffusion is approximated using a loss process $\lambda(\ell, t) f$, where $\lambda(\ell, t)$ is the *loss rate*. As an alternative it is also possible to express the loss rate as a *loss time scale* $\tau(\ell, t) = \frac{1}{\lambda(\ell,t)}$, but in this thesis we will use the former convention.

The radial diffusion system in equation (6.1) is the starting point for chapter 6 where a surrogate model of the phase space density $\hat{f}$ is built to perform Bayesian inference over the parameters of the diffusion coefficient κ and loss rate λ.

3.3.2 Current Systems & Geomagnetic Indices

As was noted earlier, the solar wind is largely deflected by the Earth's magnetic field but some particles still leak into the magnetosphere. This particle injection is governed by the interaction between the magnetic field carried by the solar wind and the Earth's magnetic field, also known as *solar wind - magnetosphere coupling*. It plays an important role in determining space weather conditions in the Earth's vicinity.

Solar wind plasma gets trapped in the Earth's magnetic field at a rate that is modulated by the solar wind - magnetosphere coupling. The drift motions of charged particles in the magnetosphere as discussed in section 3.3 lead to many current systems. The prominent current systems (pictured in figure 3.6) are 1. the ring current, 2. field aligned current, 3. tail current, and 4. magnetopause current. These current systems induce magnetic fields that interact with the Earth's magnetic field and mutate it. Weakening of the Earth's magnetic field strength due to strong ring currents leads to geomagnetic storm conditions which can have adverse impacts on orbiting satellites and ground based infrastructure.

For the purposes of space weather monitoring and forecasting, the state of the magnetosphere and geomagnetic phenomena are often represented by proxies known as geomagnetic indices. Geomagnetic indices give us the ability to summarize the state of the magnetosphere in terse framework. They are often calculated by averaging several ground based measurements of magnetic fluctuations, generally at a cadence of a few hours.

Chapter 4 gives a brief introduction to the popular geomagnetic indices and formulates *gaussian process* models for producing probabilistic one hour ahead forecasts of the Dst index. In chapter 5, we augment the Dst model from chapter 4 with *long short-term memory* (LSTM) networks and obtain five hour ahead forecasts of the Dst.

Chapter 4

Forecasting the Disturbance Storm Time Index: Gaussian Process Models

We present a methodology for generating probabilistic predictions for the *Disturbance Storm Time* (Dst) geomagnetic activity index. We focus on the *One Step Ahead* (OSA) prediction task and use the OMNI hourly resolution data to build our models. Our proposed methodology is based on the technique of *Gaussian Process Regression* (GPR). Within this framework, we develop two models; *Gaussian Process Auto-Regressive* (GP-AR) and *Gaussian Process Auto-Regressive with eXogenous inputs* (GP-ARX). We also propose a criterion to aid model selection with respect to the order of auto-regressive inputs. Finally, we test the performance of the GP-AR and GP-ARX models on a set of 63 geomagnetic storms between 1998 and 2006 and illustrate sample predictions with error bars for some of these events.

This chapter is based on the following:

Article:
M. Chandorkar, E. Camporeale, and S. Wing. Probabilistic forecasting of the disturbance storm time index: An autoregressive gaussian process approach. *Space Weather*, 15(8):1004–1019, Aug 2017. ISSN 1542-7390. doi: 10.1002/2017SW001627. URL `https://agupubs.onlinelibrary.wiley.com/doi/full/10.1002/2017SW001627`

Book Chapter:
Mandar Chandorkar and Enrico Camporeale. Chapter 9 - probabilistic forecasting of geomagnetic indices using gaussian process models. In Enrico Camporeale, Simon Wing, and Jay R. Johnson, editors, *Machine Learning Techniques for Space Weather*, pages 237 – 258. Elsevier, 2018. ISBN 978-0-12-811788-0. doi: 10.1016/B978-0-12-811788-0.00009-3. URL `http://www.sciencedirect.com/science/article/pii/B9780128117880000093`

4.1 Introduction

The magnetosphere's dynamics and its associated solar wind driver form a complex dynamical system. It is therefore instructive and greatly simplifying to use representative indices to quantify the state of geomagnetic activity.

Geomagnetic indices come in various forms. They may take continuous or discrete values and may be defined with varying time resolutions. Their values are often calculated by averaging or combining a number of readings taken by instruments, usually magnetometers, around the Earth. Each geomagnetic index is a proxy for a particular kind of phenomenon. Some popular indices are the Kp, Dst, and the AE index.

1. Kp: The Kp index is a discrete valued global geomagnetic activity index and is based on 3 hour measurements of the K-indices [Bartels and Veldkamp, 1949]. The K-index itself is a three hour long quasi-logarithmic local index of the geomagnetic activity, relative to a calm day curve for the given location.

2. AE: The Auroral Electrojet index, AE, is designed to provide a global, quantitative measure of auroral zone magnetic activity produced by enhanced ionospheric currents flowing below and within the auroral oval [Davis and Sugiura, 1966]. It is a continuous index which is calculated every hour.

3. Dst: The Disturbance Storm Time index, Dst, is a continuous valued hourly index which gives a measure of the weakening or strengthening of the Earth's equatorial magnetic field due to particle injection in the magnetosphere. Particle injection has a number of sources such as, weakening or strengthening of the ring currents and the geomagnetic storms [Dessler and Parker, 1959], near Earth cross tail current [Ganushkina et al., 2004, 2010], partial ring current [Liemohn et al., 2001], substorm current wedge [Munsami, 2000], magnetopause current, etc.

For the present study, we focus on prediction of the hourly Dst index which is a straightforward indicator of geomagnetic storms. More specifically, we focus on the one step ahead (OSA) (in this case one hour ahead) prediction of Dst because it is the simplest model towards building long term predictions of the geomagnetic response of the Earth to changing space weather conditions.

Dst OSA prediction has been the subject of several modeling efforts in the literature. One of the earliest models has been presented by Burton et al. [1975] who calculated Dst(t) as the solution of an ordinary differential equation (ODE) which expressed the rate of change of Dst(t) as a combination of decay and injection phenomena. Their model is described by the following ODE: $\frac{d}{dt}\mathrm{Dst}(t) = Q(t) - \frac{\mathrm{Dst}(t)}{\tau}$, where $Q(t)$ relates to the particle injection from the plasma sheet into the inner magnetosphere.

The Burton et al. [1975] model has proven to be very influential particularly due to its simplicity. Many subsequent works have modified the proposed ODE by proposing alternate expressions for the injection $Q(t)$ [Wang et al., 2003, O'Brien and McPherron, 2000]. More recently Ballatore and Gonzalez [2014] have tried to generate empirical estimates for the injection and decay terms in Burton's equation.

Another important empirical model used to predict Dst is the *non-linear auto-regressive moving average with exogenous inputs* (NARMAX) methodology [Billings et al., 1989, Balikhin et al., 2001, Zhu et al., 2006, 2007, Boynton et al., 2011a,b, 2013]. The NARMAX methodology builds models by constructing polynomial expansions of inputs and determines the best combinations of monomials to include in the refined model by using a criterion called the *error reduction ratio* (ERR). The parameters of the well known NARMAX OLS-ERR model are calculated by solving the ordinary least squares (OLS) problem arising from a quadratic objective function. It must be noted that the NARMAX methodology is not limited to polynomial functions, rather any set of basis function expansions can be used with it, such as radial basis functions, wavelets, etc [Wei et al., 2006, 2004]. The reader may refer to Billings [2013] for a detailed exposition of the NARMAX methodology.

Forecasting methods is based on *neural networks* have also been a popular choice for building predictive models. Researchers have employed both the standard feed forward and the more specialised recurrent architectures. Lundstedt et al. [2002] proposed an *Elman* recurrent network architecture called Lund Dst, which used the solar wind velocity, *interplanetary magnetic field* (IMF), and historical Dst data as inputs. Wing et al. [2005]

used recurrent neural networks to predict Kp. Bala et al. [2009] originally proposed a feed forward network for predicting the Kp index which used the *Boyle coupling function* [Boyle et al., 1997]. The same architecture is adapted for prediction of Dst in Bala et al. [2009], popularly known as the Rice Dst model. Pallocchia et al. [2006] proposed a neural network model called EDDA to predict Dst using only the IMF data.

Apart from the NARMAX and neural network approaches, fuzzy methods have also been applied for Dst prediction, Sharifi et al. [2006] outline the application of *local neurofuzzy* models for one hour and two hour predictions of Dst respectively. Local neurofuzzy models reduce the input space into a number of regions each with its own expert predictor. The combined model predicts Dst for a new point as a linear combination of the prediction from each expert weighted by a fuzzy score signifying the importance of each model for the provided input. For improving predictive performance of two hour Dst forecasts in Sharifi et al. [2006], the authors use *singular spectrum analysis* (SSA). Singular spectrum analysis consists of extracting orthogonal components from a lagged time series. It is equivalent to *principal component analysis* (PCA) which is quite extensively used in the machine learning community. Loskutov et al. [2001a,b] provide a good background to the theory and application of SSA to geomagnetic time series.

Although much research has been done on the prediction of the Dst index, much less has been done on probabilistic forecasting of Dst. One such work described in McPherron et al. [2013] involves identification of high speed solar wind streams using the WSA model [Wang and Sheeley, 1990], and using those predictions to construct ensembles of Dst trajectories which yield the quartiles of Dst time series.

A simple way to construct error bars on the predictions of forecasting models is by using the well known *past cast* performance, i.e. by calculating the standard deviations of the predictions generated by the model on a hold out data set. One limitation of such an approach is that the variance of the model predictions is computed once and never changes. It does not adapt according to the inputs provided to the model. This may lead to overestimation or underestimation of the uncertainty around a given prediction, depending on the prevalent geo-magnetic conditions and the data set used to calculate the *past cast* model performance.

In this chapter, we propose a technique for the probabilistic forecasting of Dst, which yields a predictive distribution as a closed form expression. Our models take as input past values of Dst, solar wind speed, and the z component of the IMF and output a Gaussian distribution with a specific mean and variance as the OSA prediction of the Dst.

We use the Gaussian process regression methodology to construct auto-regressive models for Dst and show how to perform exact inference in this framework. We further outline a methodology to perform model selection with respect to its free parameters and time histories.

The remainder of this chapter is organised into sections as follows. section 4.2 gives the reader an historical overview of *Gaussian process* (GP) models, how they are formulated, and how to perform inference with them. Section 4.3 defines the OSA prediction setting for the Dst index. Sections 4.3.1 and 4.3.2 formulate the GP-AR and GP-ARX models for making probabilistic OSA forecasts, and section 4.4 describes their training and validation process. Section 4.4 evaluates the performance of the proposed GP-AR and GP-ARX models, and section 4.5 discusses the results and offers concluding remarks.

4.2 Methodology: Gaussian Process

Gaussian processes first appeared in machine learning research in Neal [1996], as the limiting case of Bayesian inference performed on neural networks with infinitely many neurons in the hidden layers. Although their inception in the machine learning community is recent, their origins can be traced back to the geo-statistics research community where they are known as *Kriging* methods [Krige, 1951]. In pure mathematics, Gaussian processes have been studied extensively and their existence was first proven by Kolmogorov's extension theorem [Tao, 2011]. The reader is referred to Rasmussen and Williams [2005] for an in-depth treatment of Gaussian processes in machine learning.

Let us assume that we want to model a process in which a scalar quantity y is specified as $y = f(\mathbf{x}) + \epsilon$ where $f(.) : \mathbb{R}^d \rightarrow \mathbb{R}$ is an unknown scalar function of a multidimensional input vector $\mathbf{x} \in \mathbb{R}^d$, d is the dimensionality of the input space, and $\epsilon \sim \mathcal{N}(0, \sigma^2)$ is zero mean Gaussian noise with variance σ^2.

A set of labeled data points $(\mathbf{x}_i, y_i); i = 1 \cdots N$ can be conveniently expressed by a $N \times d$ data matrix $\mathbf{X}$ and a $N \times 1$ response vector $\mathbf{y}$, as shown in equations (4.1) and (4.2).

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_N^T \end{pmatrix}_{N \times d} \tag{4.1}$$

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}_{N \times 1} \tag{4.2}$$

Our task is to infer the values of the unknown function $f(.)$ based on the inputs $\mathbf{X}$ and the noisy observations $\mathbf{y}$. We now assume that the joint distribution of $f(\mathbf{x}_i), i = 1 \cdots N$ is a multivariate Gaussian as shown in equations (4.3) to (4.5).

$$\mathbf{f} = \begin{pmatrix} f(\mathbf{x}_1) \\ f(\mathbf{x}_2) \\ \vdots \\ f(\mathbf{x}_N) \end{pmatrix} \tag{4.3}$$

$$\mathbf{f}|\mathbf{x}_1, \cdots, \mathbf{x}_N \sim \mathcal{N}\left(\mu, \mathbf{\Lambda}\right) \tag{4.4}$$

$$p(\mathbf{f} \mid \mathbf{x}_1, \cdots, \mathbf{x}_N) = \frac{1}{(2\pi)^{n/2} \det(\mathbf{\Lambda})^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{f} - \mu)^T \mathbf{\Lambda}^{-1} (\mathbf{f} - \mu)\right) \tag{4.5}$$

Here $\mathbf{f}$ is a $N \times 1$ vector consisting of the values $f(\mathbf{x}_i), i = 1 \cdots N$. In equation (4.4), $\mathbf{f}|\mathbf{x}_1, \cdots, \mathbf{x}_N$ denotes the conditional distribution of $\mathbf{f}$ with respect to the input data (i.e., $\mathbf{X}$) and $\mathcal{N}\left(\mu, \mathbf{\Lambda}\right)$ represents a multivariate Gaussian distribution with mean vector μ and covariance matrix $\mathbf{\Lambda}$. The probability density function of this distribution $p(\mathbf{f} \mid \mathbf{x}_1, \cdots, \mathbf{x}_N)$ is therefore given by equation (4.5).

From equation (4.5), one can observe that in order to uniquely define the distribution of the process, it is required to specify μ and $\mathbf{\Lambda}$. For this probability density to be valid, there are further requirements imposed on $\mathbf{\Lambda}$:

1. Symmetry: $\mathbf{\Lambda}_{ij} = \mathbf{\Lambda}_{ji} \; \forall i, j \in 1, \cdots, N$
2. Positive Semi-definiteness: $\mathbf{z}^T \mathbf{\Lambda} \mathbf{z} \geq 0 \; \forall \mathbf{z} \in \mathbb{R}^N$

Inspecting the individual elements of μ and $\mathbf{\Lambda}$, we realise that they take the following form.

$$\mu_i = \mathbb{E}[f(\mathbf{x}_i)] := m(\mathbf{x}_i) \tag{4.6}$$

$$\Lambda_{ij} = \mathbb{E}[(f(\mathbf{x}_i) - \mu_i)(f(\mathbf{x}_j) - \mu_j)] := K(\mathbf{x}_i, \mathbf{x}_j) \tag{4.7}$$

Here $\mathbb{E}$ denotes the expectation (average). The elements of μ and $\mathbf{\Lambda}$ are expressed as functions $m(\mathbf{x}_i)$ and $K(\mathbf{x}_i, \mathbf{x}_j)$ of the inputs $\mathbf{x}_i, \; \mathbf{x}_j$. Specifying the functions $m(\mathbf{x})$ and $K(\mathbf{x}, \mathbf{x}')$ completely specifies each element of μ and $\mathbf{\Lambda}$ and subsequently the finite dimensional distribution of $\mathbf{f}|\mathbf{x}_1, \cdots, \mathbf{x}_N$. In most practical applications of Gaussian processes, the mean function is often defined as $m(\mathbf{x}) = 0$, which is not unreasonable if the data is standardised to have zero mean. Gaussian processes are represented in machine learning literature using the following notation:

$$f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), K(\mathbf{x}, \mathbf{x}')) \tag{4.8}$$

4.2.1 Inference and Predictions

Our aim is to infer the function $f(\mathbf{x})$ from the noisy training data and generate predictions $f(\mathbf{x}_i^*)$ for a set of test points $\mathbf{x}_i^* : \forall i \in 1, \cdots, M$. We define $\mathbf{X}^*$ as the test data matrix whose rows are formed by $\mathbf{x}_i^*$ as shown in equation (4.9).

$$\mathbf{X}_* = \begin{pmatrix} (\mathbf{x}_1^*)^T \\ (\mathbf{x}_2^*)^T \\ \vdots \\ (\mathbf{x}_M^*)^T \end{pmatrix}_{M \times d} \tag{4.9}$$

Using the multivariate Gaussian distribution in equation (4.5), we can construct the joint distribution of $f(\mathbf{x})$ over the training and test points. The vector of training and test outputs $\begin{pmatrix} \mathbf{y} \\ \mathbf{f}_* \end{pmatrix}$ is of dimension $(N + M) \times 1$

and is constructed by appending the test set predictions $\mathbf{f}_*$ to the observed noisy measurements $\mathbf{y}$.

$$\mathbf{f}_* = \begin{pmatrix} f(\mathbf{x}^*_\mathbf{1}) \\ f(\mathbf{x}^*_\mathbf{2}) \\ \vdots \\ f(\mathbf{x}^*_\mathbf{M}) \end{pmatrix}_{M \times 1} \tag{4.10}$$

$$\begin{pmatrix} \mathbf{y} \\ \mathbf{f}_* \end{pmatrix} \mid \ \mathbf{X}, \mathbf{X}_* \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \mathbf{K} + \sigma^2 \mathbf{I} & \mathbf{K}_* \\ \mathbf{K}_*^T & \mathbf{K}_{**} \end{bmatrix}\right) \tag{4.11}$$

Since we have noisy measurements of f over the training data, we add the noise variance σ^2 to the variance of f as shown in equation (4.11). The block matrix components of the $(N + M) \times (N + M)$ covariance matrix have the following structure.

1. $\mathbf{I}$: The $N \times N$ identity matrix.
2. $\mathbf{K} = [K(\mathbf{x}_i, \mathbf{x}_j)], \ i, j \in 1, \cdots, N$: Kernel matrix constructed from all couples obtained from the training data.
3. $\mathbf{K}_* = [K(\mathbf{x}_i, \mathbf{x}^*_j)], \ i \in 1, \cdots, N; j \in 1, \cdots, M$: Cross kernel matrix constructed from all couples between training and test data points.
4. $\mathbf{K}_{**} = [K(\mathbf{x}^*_i, \mathbf{x}^*_j)], \ i, j \in 1, \cdots, M$: Kernel matrix constructed from all couples obtained from the test data.

Using the multivariate normal distribution in equation (4.11), probabilistic predictions f_* can be generated by constructing the conditional distribution $\mathbf{f}_* | \mathbf{X}, \mathbf{y}, \mathbf{X}_*$. Since the original distribution of $\begin{pmatrix} \mathbf{y} \\ \mathbf{f}_* \end{pmatrix} \mid \ \mathbf{X}, \mathbf{X}_*$ is a multivariate Gaussian, conditioning on a subset of elements $\mathbf{y}$ yields another Gaussian distribution whose mean and covariance can be calculated exactly, as in equation (4.12) [Rasmussen and Williams, 2005].

$$\mathbf{f}_* | \mathbf{X}, \mathbf{y}, \mathbf{X}_* \sim \mathcal{N}(\bar{\mathbf{f}}_*, \Sigma_*), \tag{4.12}$$

where

$$\bar{\mathbf{f}}_* = \mathbf{K}_*^T [\mathbf{K} + \sigma^2 \mathbf{I}]^{-1} \mathbf{y} \tag{4.13}$$

$$\Sigma_* = \mathbf{K}_{**} - \mathbf{K}_*^T \left(\mathbf{K} + \sigma^2 \mathbf{I}\right)^{-1} \mathbf{K}_* \tag{4.14}$$

The practical implementation of GP models requires the inversion of the training data kernel matrix $[\mathbf{K} + \sigma^2\mathbf{I}]^{-1}$ to calculate the parameters of the predictive distribution $\mathbf{f}_*|\mathbf{X}, \mathbf{y}, \mathbf{X}_*$. The computational complexity of this inference is dominated by the linear problem in equation (4.13), which can be solved via Cholesky decomposition, with a time complexity of $O(N^3)$, where N is the number of data points.

The distribution of $\mathbf{f}_*|\mathbf{X}, \mathbf{y}, \mathbf{X}_*$ is known in Bayesian analysis as the *posterior predictive distribution*. This illustrates a key difference between Gaussian processes and other regression models such as *neural networks*, *linear models* and *support vector machines*: a GP model does not generate point predictions for new data but outputs a predictive distribution for the quantity sought, thus allowing to construct error bars on the predictions. This property which is common to Bayesian models makes them very appealing for Space Weather forecasting applications.

The central design issue in applying GP models is the choice of the function $K(\mathbf{x}, \mathbf{x}')$. The same constraints that apply to $\mathbf{\Lambda}$ also apply to the function K. In machine learning, these symmetric positive definite functions of two variables are known as covariance functions or *kernels*. Kernel based methods are applied extensively in data analysis [Scholkopf and Smola, 2001, Hofmann et al., 2008] i.e. regression, clustering, classification, and density estimation [Girolami, 2002].

Table 4.1: Popular Kernel functions used in GPR models

Name	**Expression**	**Hyperparameters**
Radial Basis Function (RBF)	$\frac{1}{2}\exp(-\|\|\mathbf{x} - \mathbf{y}\|\|^2/l^2)$	$l \in \mathbb{R}$
Polynomial	$(\mathbf{x} \cdot \mathbf{y} + b)^d$	$b \in \mathbb{R}, d \in \mathbb{N}$
Laplacian	$\exp(-\|\|\mathbf{x} - \mathbf{y}\|\|_1/\theta)$	$\theta \in \mathbb{R}^+$
Student's T	$1/(1 + \|\|\mathbf{x} - \mathbf{y}\|\|_2^d)$	$d \in \mathbb{R}^+$
Maximum Likelihood Perceptron (MLP)	$\sin^{-1}\left(\frac{w\mathbf{x}\cdot\mathbf{y}+b}{\sqrt{w\mathbf{x}\cdot\mathbf{x}+b+1}\sqrt{w\mathbf{y}\cdot\mathbf{y}+b+1}}\right)$	$w, b \in \mathbb{R}^+$

4.2.2 Kernel Functions

For the success of a GP model an appropriate choice of kernel function is paramount. The symmetry and positive semi-definiteness of GP kernels implies that they represent inner-products between some basis function representation of the data. The interested reader is suggested to refer to

Berlinet and Thomas-Agnan [2004], Scholkopf and Smola [2001], Hofmann et al. [2008] for a thorough treatment of kernel functions and the rich theory behind them. Some common kernel functions used in machine learning are listed in table 4.1.

The quantities l in the RBF, and b and d in the polynomial kernel are known as *hyper-parameters*. Hyper-parameters give flexibility to a particular kernel structure, for example $d = 1, 2, 3, \dots$ in the polynomial kernel represents linear, quadratic, cubic and higher order polynomials respectively. The process of assigning values to the hyper-parameters is crucial in the model building process and is known as *model selection*.

4.2.3 Model Selection

Given a GP model with a kernel function K_θ, the problem of model selection consists of finding appropriate values for the kernel hyper-parameters $\theta = (\theta_1, \theta_2, \cdots, \theta_i)$. In order to assign a value to θ, we must define an objective function which represents our confidence that the GP model built from a particular value of θ is the best performing model. Since GP models encode assumptions about the probability distribution of the output data $\mathbf{y}$ given inputs $\mathbf{X}$, it is natural to use the negative log-likelihood of the training data as a model selection criterion.

$$\begin{aligned}
Q(\theta) &= -\log(p(\mathbf{y}|\mathbf{X}, K_\theta)) \\
&= -\frac{1}{2}\mathbf{y} \cdot (\mathbf{K}_\theta + \sigma^2\mathbf{I})^{-1}\mathbf{y} - \frac{1}{2}|\mathbf{K}_\theta + \sigma^2\mathbf{I}| - \frac{N}{2}\log(2\pi) \\
\mathbf{K}_\theta &= [K_\theta(\mathbf{x}_i, \mathbf{x}_j)]_{N\times N}
\end{aligned}$$

The model selection problem can now be expressed as the minimization problem shown below.

$$\theta^* = arg\min_\theta \; Q(\theta)$$

The objective function $Q(\theta)$ in the general case can have multiple local minima, and evaluating the value of $Q(.)$ at any given θ requires inversion of the matrix $\mathbf{K}_\theta + \sigma^2\mathbf{I}$ which has a time complexity $O(N^3)$ as noted above. In the interest of saving computational cost, one cannot use exhaustive search through the domain of the hyper-parameters to inform our choice for θ. Some of the techniques used for model selection in the context of GPR include.

1. *Grid search* (GS): This method constructs a grid of values for θ by taking cartesian product of one dimensional grids for each θ_i. It evaluates $Q(.)$ at each grid point and chooses the configuration which yields the lowest value of $Q(.)$.

2. *Coupled simulated annealing* (CSA): Introduced in Xavier-De-Souza et al. [2010], CSA follows the same procedure as grid search, but after the evaluation of $Q(.)$ on the grid, each grid point is iteratively mutated in a random walk fashion. This mutation is accepted or rejected according to the new value of $Q(.)$ as well as its value on the other grid points. This procedure is iterated until some stop criterion is reached.

3. *Maximum likelihood* (ML): This technique outlined in Rasmussen and Williams [2005] is a form of gradient descent. It involves starting with an initial guess for θ and iteratively improving it by calculating the gradient of $Q(.)$ with respect to θ. Although this method seems intuitive, it introduces an extra computational cost of calculating the gradient of $Q(\theta)$ with respect to each θ_i in every iteration, and applying this method can sometimes lead to overfitting of the GPR model to the training data [Rasmussen and Williams, 2005, ch. 5,sec. 5.2].

4.3 One Step Ahead Prediction

In equations (4.15) to (4.17), we outline a Gaussian process formulation for the OSA prediction of Dst. A vector of features $\mathbf{x}_{t-1}$ is used as input to an unknown function $f(\mathbf{x}_{t-1})$.

The features $\mathbf{x}_{t-1}$ can be any collection of quantities in the hourly resolution OMNI data set. Generally, $\mathbf{x}_{t-1}$ are time histories of Dst and other important variables such as the plasma pressure $p(t)$, solar wind speed $V(t)$, and the z component of the IMF $B_z(t)$.

$$\mathrm{Dst}(t) = f(\mathbf{x}_{t-1}) + \epsilon \tag{4.15}$$

$$\epsilon \sim \mathcal{N}(0, \sigma^2) \tag{4.16}$$

$$f(x_t) \sim \mathcal{GP}(m(\mathbf{x}_t), K_{\mathrm{osa}}(\mathbf{x}_t, \mathbf{x}_s)) \tag{4.17}$$

We consider two choices for the input features $\mathbf{x}_{t-1}$ leading to two variants of Gaussian process regression for Dst time series prediction.

4.3.1 Gaussian Process Auto-Regressive (GP-AR)

The simplest auto-regressive models for OSA prediction of Dst are those that use only the history of Dst to construct input features for model training. The input features $\mathbf{x}_{t-1}$ at each time step are the history of Dst(t) until a time lag of p hours.

$$\mathbf{x}_{t-1} = (\mathrm{Dst}(t-1), \cdots, \mathrm{Dst}(t-p+1)) \tag{4.18}$$

This description of the GP-AR model in equations (4.15) to (4.18), as a Gaussian process $Dst(t)$ whose input space $\mathbf{x}(t)$ comprises previous realizations of itself, has been discussed in time series [Roberts et al., 2013], control systems [Kocijan, 2015], and machine learning literature [Wang et al., 2006, 2007]. The reader may refer to appendix A for detailed notes regarding GP time series models, some nuances of the GP-AR formulation,and its relationship to auto-regressive time series models.

4.3.2 Gaussian Process Auto-Regressive with eXogenous inputs (GP-ARX)

Auto-regressive models can be augmented by including exogenous quantities in the inputs $\mathbf{x}_{t-1}$ at each time step, in order to improve predictive accuracy. While modeling Dst using the OMNI data, one must choose which solar wind quantities to include in the exogenous inputs of the predictive model. This choice is not straight forward and eventually requires a compromise between including important solar wind quantities and keeping the input space manageable in the interest of simplicity.

Choice of Solar Wind Inputs

Dst gives a measure of ring currents, which are modulated by plasma sheet particle injections into the inner magnetosphere during sub-storms. Studies have shown that the substorm occurrence rate increases with solar wind velocity i.e. high speed streams [Kissinger et al., 2011, Newell et al., 2016]. Prolonged southward pointing B_z is needed for sub-storms to occur [McPherron et al., 1986]. An increase in the solar wind electric field, $V_{\mathrm{sw}} B_z$, can increase the dawn-dusk electric field in the magnetotail, which in turn determines the amount of plasma sheet particle that move into the inner magnetosphere [Friedel et al., 2001].

Apart from V and B_z, other quantities which have been shown to correlate with geomagnetic activity are solar wind dynamic pressure P, clock angle $\tan\theta = \frac{B_y}{B_z}$, Akasofu ϵ [Pudovkin and Semenov, 1986], and solar wind magnetosphere coupling functions [Spencer et al., 2011].

Although solar wind magnetospheric coupling functions correlate with geomagnetic indices, they are expressed in terms of V and B_z, and hence we do not include them as explicit inputs to the model. Gaussian process models derive their strength from automatic feature construction achieved by the covariance functions (interested readers may refer to Rasmussen and Williams [2005, ch. 6 & 7]). As long as coupling functions can be approximated in the eigenspace of the covariance function we need not make them explicit in the input features.

Therefore, our exogenous parameters consist of solar wind velocity V_{sw} and IMF B_z. In this model, we choose distinct time lags p, p_v and p_b for Dst, V and B_z respectively.

$$\begin{aligned}\mathbf{x}_{t-1} = (&\mathrm{Dst}(t-1), \cdots, \mathrm{Dst}(t-p+1),\\ &V_{\mathrm{sw}}(t-1), \cdots, V_{\mathrm{sw}}(t-p_v+1),\\ &B_z(t-1), \cdots, B_z(t-p_b+1))\end{aligned}$$

It is an important question as to how unaccounted inputs such as solar wind dynamic pressure P and clock angle θ affect the structure of the GP-ARX model. From a model selection perspective, these unaccounted inputs should lead to higher values of the noise covariance. In the specific case of solar wind dynamic pressure, it is calculated as a product of the plasma density and the solar wind speed, making it highly correlated with the latter. As a result, the GP-ARX model can infer a large portion of the information content from the solar wind speed itself. With respect to the clock angle, it must be noted that coupling functions such as the Akasofu ϵ generally contain powers of $\sin\theta$, bounding the effect of clock angle to an absolute magnitude of 1. Hence we do not expect these unaccounted inputs to greatly improve the predictive capabilities of the GP-ARX model.

4.3.3 Choice of Mean Function

Mean functions in GPR models encode trends in the data. They are the baseline predictions the model falls back to in case the training and test data have little correlation as predicted by the kernel function. If there is no prior knowledge about the function to be approximated, Rasmussen and Williams [2005] state that it is perfectly reasonable to choose $m(\mathbf{x} = 0)$ as the mean function, as long as the target values are normalised. In the case of the Dst time series, it is known that the so called *persistence model* $\hat{\mathrm{D}}\mathrm{st}(t) = \mathrm{Dst}(t-1)$ has high OSA correlation with Dst. Due to its simplicity, we chose the persistence model as the prior mean function in our OSA Dst models.

The persistence model can be described as Markov prediction mechanism. When it is chosen as the mean function of the GP-AR and GP-ARX model, the prior probability distribution of $\mathrm{Dst}(t)$ is Gaussian with a strong Markovian behavior

$$P(\mathrm{Dst}(t)|\mathbf{x}_t) \sim \mathcal{N}(\mathrm{Dst}(t-1), \sqrt{K_{\mathrm{osa}}(\mathbf{x}_t, \mathbf{x}_t)}) \ ,$$

but the posterior predictive distribution of $\mathrm{Dst}(t)$ conditional on the model training data (given in equation (4.13)) is non-Markovian due its dependence on the term denoted by $\mathbf{K}_*$ which contains kernel values computed between the test data and training data features. Thus, the GP-AR and GP-ARX models when used conditional on training data are non-Markovian predictive models.

4.3.4 Choice of Kernel

In this study, we constructed Gaussian Process regression models with a combination of the *maximum likelihood perceptron* kernel, and *student's T* kernel as shown in equation (4.19). The maximum likelihood perceptron kernel is the Gaussian process equivalent of a single hidden layer feed-forward neural network model as demonstrated in Neal [1996].

$$K_{\text{osa}}(\mathbf{x},\mathbf{y}) = K_{\text{mlp}}(\mathbf{x},\mathbf{y}) + K_{\text{st}}(\mathbf{x},\mathbf{y}) \tag{4.19}$$

$$K_{\text{mlp}}(\mathbf{x},\mathbf{y}) = \sin^{-1}\left(\frac{w\mathbf{x}\cdot\mathbf{y}+b}{\sqrt{w\mathbf{x}\cdot\mathbf{x}+b+1}\sqrt{w\mathbf{y}\cdot\mathbf{y}+b+1}}\right) \tag{4.20}$$

$$K_{\text{st}}(\mathbf{x},\mathbf{y}) = \frac{1}{1+||\mathbf{x}-\mathbf{y}||_2^d} \tag{4.21}$$

4.4 Experiments

Table 4.2: Settings of model selection procedures

Procedure	**Grid Size**	**Step**	**Max Iterations**
Grid Search	10	0.2	NA
Coupled Simulated Annealing	4	0.2	30
Maximum likelihood	NA	0.2	150

Training

We chose data from the time periods 00 : 00 January 3 2010 to 23 : 00 January 23 2010 and 20 : 00 August 5 2011 to 22 : 00 August 6 2011 for training the GP-AR and GP-ARX models. The first training data period consists of ambient fluctuations of Dst while the second contains a geomagnetic storm.

As discussed in section 4.2.1, the computational complexity of calculating the GP predictive distribution is $O(N^3)$. This can limit the size of the covariance matrix constructed from the training data. Note that this computational overhead is paid for every unique assignment to the model's hyper-parameters. However, our chosen training set has a size of 243 which is still very much below the computational limits of the method and in our case solving equation (4.13) on a laptop computer takes less than a second for the training set considered in our analysis.

Selection

In order to find appropriate values of the hyper-parameters of the chosen kernel K_{osa}, we used the grid search, coupled simulated annealing, and maximum likelihood methods. We fixed the parameters d and σ^2 of K_{st} and model noise to values 0.01 and 0.2 respectively. The remaining parameters w and b are kept free to be determined by model selection. Table 4.2 summarises the settings used to run each model selection procedure.

Validation

Apart from selecting the kernel parameters, one also needs to choose appropriate values for the auto-regressive orders p in the case of GP-AR and p, p_v, p_b in the case of GP-ARX. For this purpose we used a set of 24 storm events listed in table 4.4 and for every assignment of values to the model order, we performed model selection with the routines in table 4.2 and recorded the performance on this validation set.

For measuring performance of model instances on the validation set storm events, the following metrics were calculated.

1. The mean absolute error

$$\mathrm{MAE} = \sum_{t=1}^{n} \left| (\mathrm{Dst}(t) - \hat{\mathrm{D}}\mathrm{st}(t)) \right| / n \ . \tag{4.22}$$

2. The root mean square error

$$\mathrm{RMSE} = \sqrt{\sum_{t=1}^{n} (\mathrm{Dst}(t) - \hat{\mathrm{D}}\mathrm{st}(t))^2 / n} \ . \tag{4.23}$$

3. The Pearson correlation coefficient between the predicted and actual value of Dst

$$\mathrm{CC} = \mathrm{Cov}(\mathrm{Dst}, \hat{\mathrm{D}}\mathrm{st}) / \sqrt{\mathrm{Var}(\mathrm{Dst})\,\mathrm{Var}(\hat{\mathrm{D}}\mathrm{st})} \ . \tag{4.24}$$

In the case of GP-AR, we let the model order p vary from 5 to 12. For GP-ARX we let the total model order $p_t = p + p_v + p_b$ vary from 3 to 12, and for each p_t evaluated every possible combination of p, p_v and p_b such that $p_t = p + p_v + p_b$ and $p, p_v, p_b > 0$.

Evaluation

After selecting the best performing GP-AR and GP-ARX models in the validation phase, we tested and compared the performance of these models with the predictions generated from the persistence model on a set of 63 storm events occurring between 1998 and 2006 as given in table 4.5, which is the same list of storm events as used in Ji et al. [2012].

Table 4.3: Evaluation results for models on storm events listed in table 4.5

Model	MAE	RMSE	CC
GP-ARX	7.219 nT	11.88 nT	0.972
GP-AR	8.37 nT	14.04 nT	0.963
Persistence	9.182 nT	14.94 nT	0.957

4.5 Results

Figures 4.1 and 4.2 show how the mean absolute error and the coefficient of correlation calculated on the validation set storm events from table 4.4, vary with increasing model order for GP-AR and GP-ARX. The results are represented as box and whisker plots, in which a rectangle is drawn to represent the first and third quartiles, with a horizontal line inside to indicate the median value. Outlying points are shown as dots while the whiskers indicate the smallest and largest non-outliers. In both cases, the predictive performance first improves and then stagnates or worsens with increasing model order.

Figures 4.3 and 4.4 break down the results for GP-ARX by the model selection routine used. Apart from the general trend observed in figures 4.1 and 4.2, we also observe that grid search and coupled simulated annealing give superior performance when compared with gradient-based maximum likelihood.

From the validation results, we choose the model order which yields the best RMSE performance. For GP-AR it is $p_t = 6$ while for GP-ARX it is $p = 6, p_v = 1, p_b = 3$.

After choosing the best performing GP-AR and GP-ARX models, we calculated their performance on the test set of table 4.5. The results of

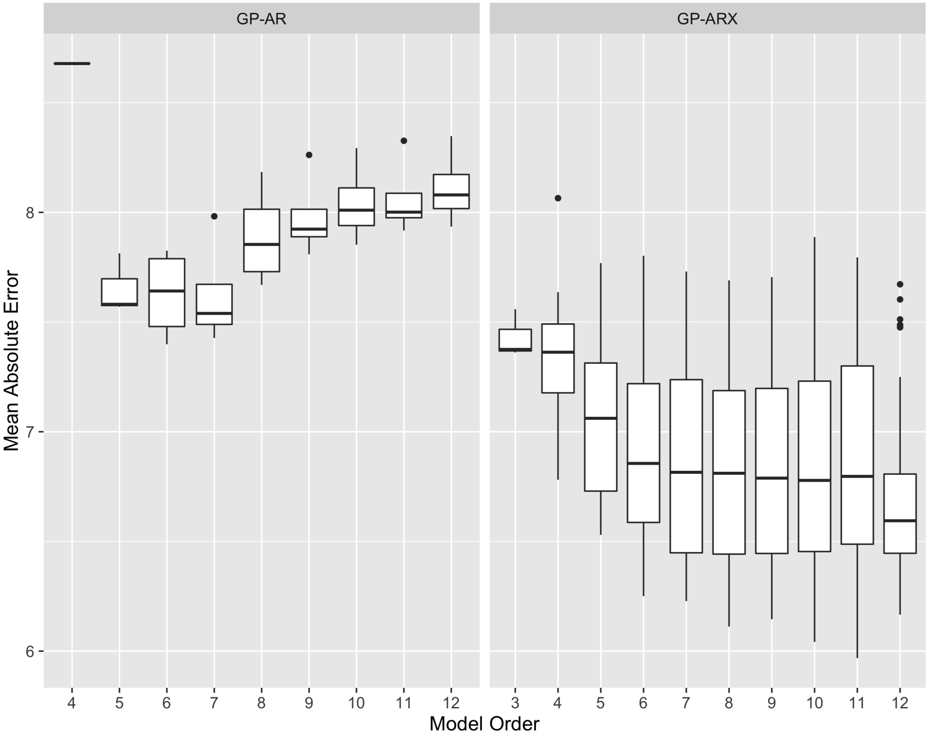


Figure 4.1: Mean Absolute Error (nT) on validation set storms vs model order for GP-AR and GP-ARX.
Key: Rectangle borders represent the first and third quartiles, with a horizontal line inside to indicate the median value. Outlying points are shown as dots and whiskers indicate the smallest and largest non-outliers

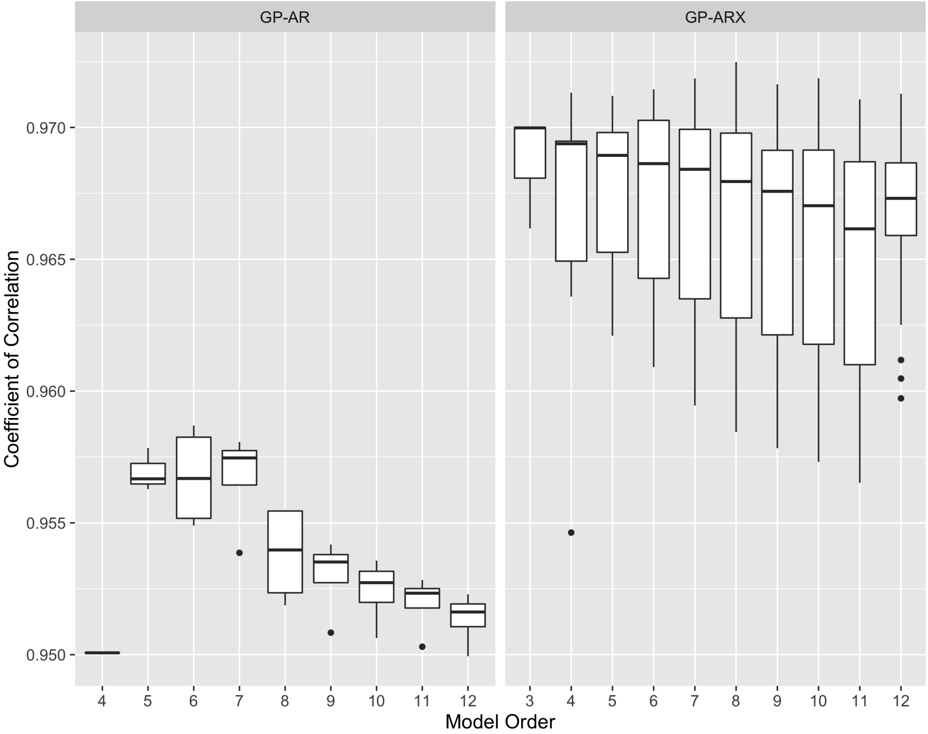


Figure 4.2: Coefficient of Correlation on validation set storms vs model order for GP-AR and GP-ARX

Key: Rectangle borders represent the first and third quartiles, with a horizontal line inside to indicate the median value. Outlying points are shown as dots and whiskers indicate the smallest and largest non-outliers

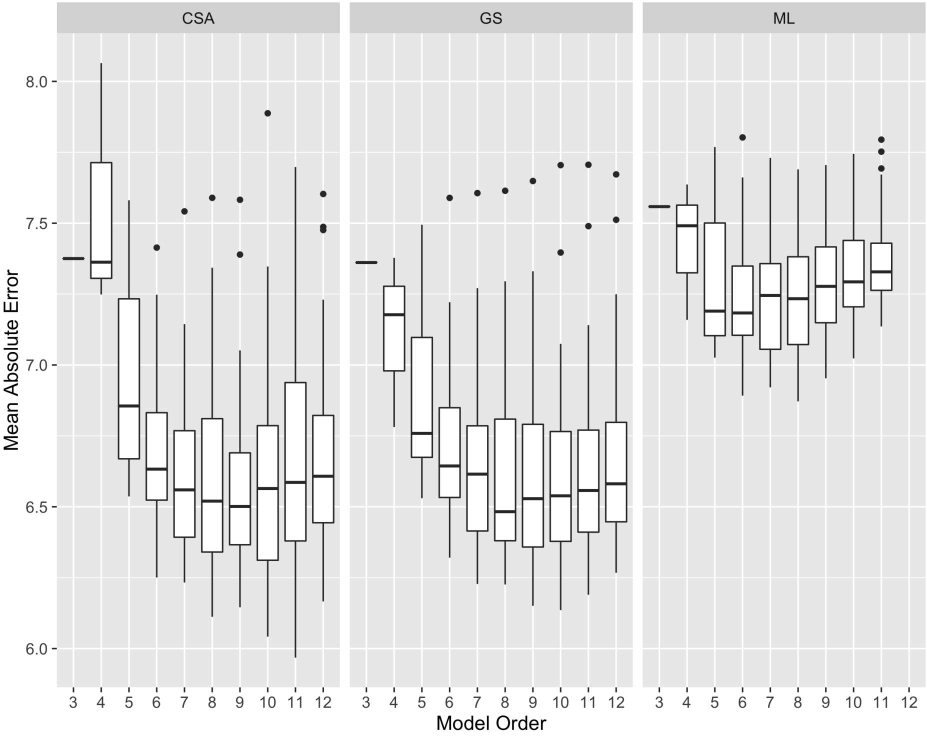


Figure 4.3: Mean Absolute Error (nT) on validation set storms vs model order for GP-AR and GP-ARX for the CSA, GS, and ML model selection routines

Key: Rectangle borders represent the first and third quartiles, with a horizontal line inside to indicate the median value. Outlying points are shown as dots and whiskers indicate the smallest and largest non-outliers

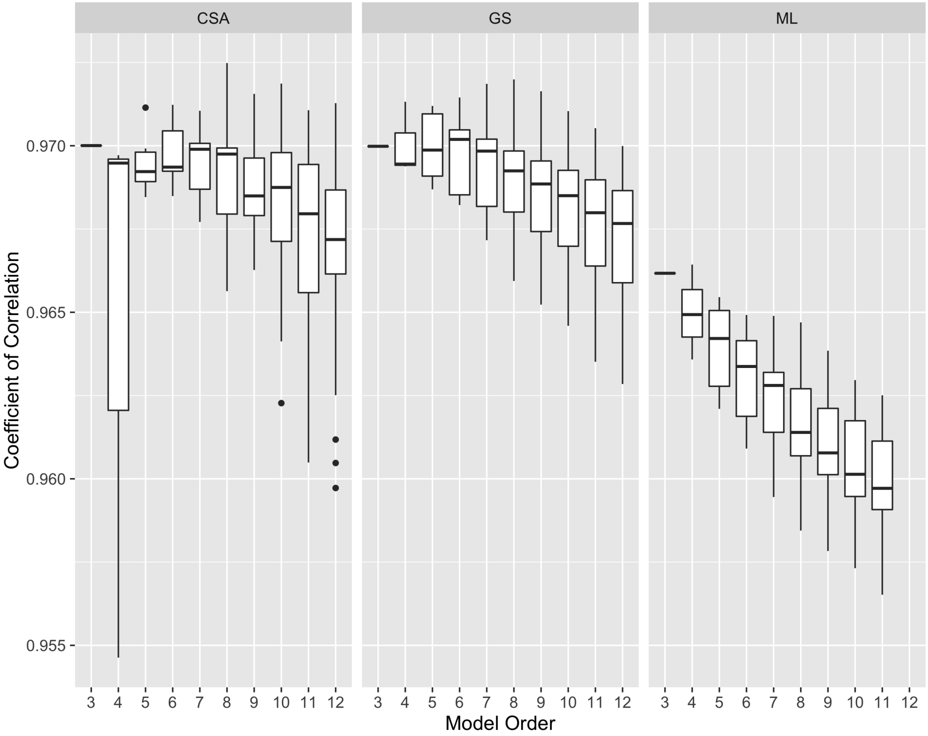


Figure 4.4: Coefficient of Correlation on validation set storms vs model order for GP-AR and GP-ARX for the CSA, GS, and ML model selection routines
Key: Rectangle borders represent the first and third quartiles, with a horizontal line inside to indicate the median value. Outlying points are shown as dots and whiskers indicate the smallest and largest non-outliers

these model evaluations are summarised in table 4.3. The GP-AR and GP-ARX models improve upon the performance of the persistence model.

Figures 4.5 to 4.7 show OSA predictions of the GP-ARX model with $\pm\sigma$ error bars for three storm events in the time period between 1998 and 2003. The GP-ARX model gives accurate predictions along with plausible error bars around its mean predictions.

4.6 Conclusions

In this chapter, we described a flexible and expressive methodology for generating probabilistic forecasts of the Dst index. We proposed two Gaussian process auto-regressive models, GP-ARX and GP-AR, to generate hourly predictions and their associated error bars. We also describe how to carry out model selection and validation of GP-AR and GP-ARX models.

Our results can be summarised as follows.

1. The persistence model plays an important role in the model building and evaluation process in the context of OSA prediction of the Dst index. Although it is not a robust predictor for the onset of intense geomagnetic storms, the persistence model performs well on classical error metrics such as RMSE and CC. From the considerations above, it is quite evident that classical performance metrics are not adequate for model evaluation. Nevertheless in space weather literature, metrics like RMSE are very commonly used to compare predictive performance of models. Although not the research focus of this study, we note that there exists a need for the formulation of more informative performance metrics for the measurement of the predictive performance of geomagnetic predictive models.

2. Gaussian process AR and ARX models give encouraging benefits in OSA prediction. Leveraging the strengths of the Bayesian approach, they are able to learn robust predictors from data. If one considers the size of the data used in our study, one can appreciate that the models presented here need relatively small training and validations sets: the training set contains 243 instances, while the validation set contains 782 instances.

3. Because GP models generate predictive distributions and not just point predictions, they lend themselves to the requirements of space weather prediction very well.
4. The Gaussian process regression framework described in this study can also be extended to multiple hour ahead prediction of Dst, which is currently a work in progress.

Table 4.4: Storm events used for model selection of GP-AR and GP-ARX

Event Id	**Start Date**	**Start Hour**	**End Date**	**End Hour**	**Storm Peak**
1	1995/03/26	05:00	1995/03/26	23:00	$-107\,\mathrm{nT}$
2	1995/04/07	13:00	1995/04/09	09:00	$-149\,\mathrm{nT}$
3	1995/09/27	01:00	1995/09/28	04:00	$-108\,\mathrm{nT}$
4	1995/10/18	13:00	1995/10/19	14:00	$-127\,\mathrm{nT}$
5	1996/10/22	22:00	1996/10/23	11:00	$-105\,\mathrm{nT}$
6	1997/04/21	10:00	1997/04/22	09:00	$-107\,\mathrm{nT}$
7	1997/05/15	03:00	1997/05/16	00:00	$-115\,\mathrm{nT}$
8	1997/10/10	18:00	1997/10/11	19:00	$-130\,\mathrm{nT}$
9	1997/11/07	00:00	1997/11/07	18:00	$-110\,\mathrm{nT}$
10	1997/11/22	21:00	1997/11/24	04:00	$-108\,\mathrm{nT}$
11	2005/06/12	17:00	2005/06/13	19:00	$-106\,\mathrm{nT}$
12	2005/08/31	12:00	2005/09/01	12:00	$-122\,\mathrm{nT}$
13	2006/12/14	21:00	2006/12/16	03:00	$-162\,\mathrm{nT}$
14	2011/09/26	14:00	2011/09/27	12:00	$-101\,\mathrm{nT}$
15	2011/10/24	20:00	2011/10/25	14:00	$-132\,\mathrm{nT}$
16	2012/03/08	12:00	2012/03/10	16:00	$-131\,\mathrm{nT}$
17	2012/04/23	11:00	2012/04/24	13:00	$-108\,\mathrm{nT}$
18	2012/07/15	01:00	2012/07/16	23:00	$-127\,\mathrm{nT}$
19	2012/09/30	13:00	2012/10/01	18:00	$-119\,\mathrm{nT}$
20	2012/10/08	02:00	2012/10/09	17:00	$-105\,\mathrm{nT}$
21	2012/11/13	18:00	2012/11/14	18:00	$-108\,\mathrm{nT}$
22	2013/03/17	07:00	2013/03/18	10:00	$-132\,\mathrm{nT}$
23	2013/05/31	18:00	2013/06/01	20:00	$-119\,\mathrm{nT}$
24	2014/02/18	15:00	2014/02/19	16:00	$-112\,\mathrm{nT}$

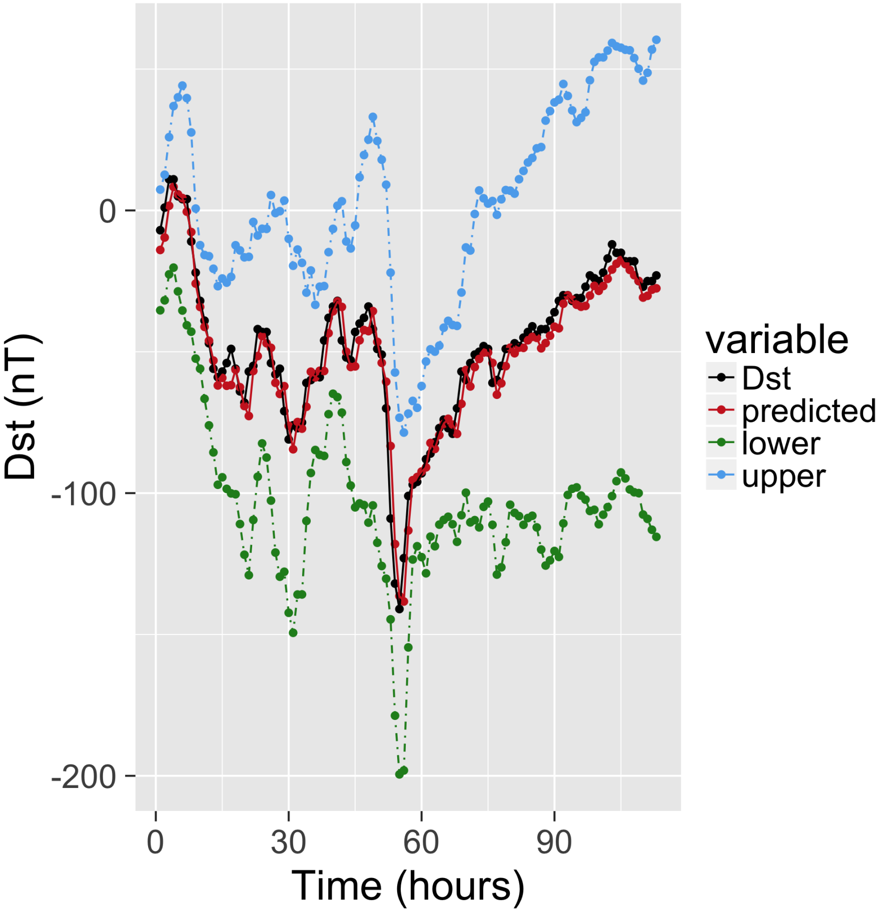


Figure 4.5: OSA Predictions with ±σ error bars for event: 2003/06/17 to 2003/06/19

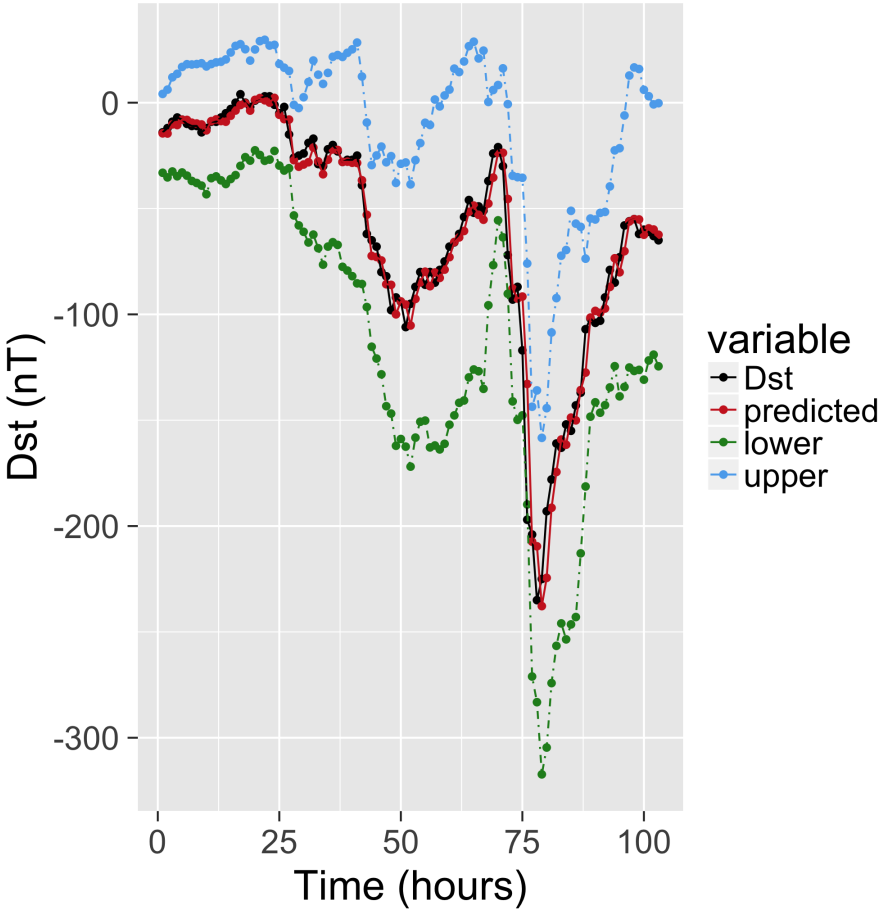


Figure 4.6: OSA Predictions with ±σ error bars for event: 2012/03/08 to 2012/03/10

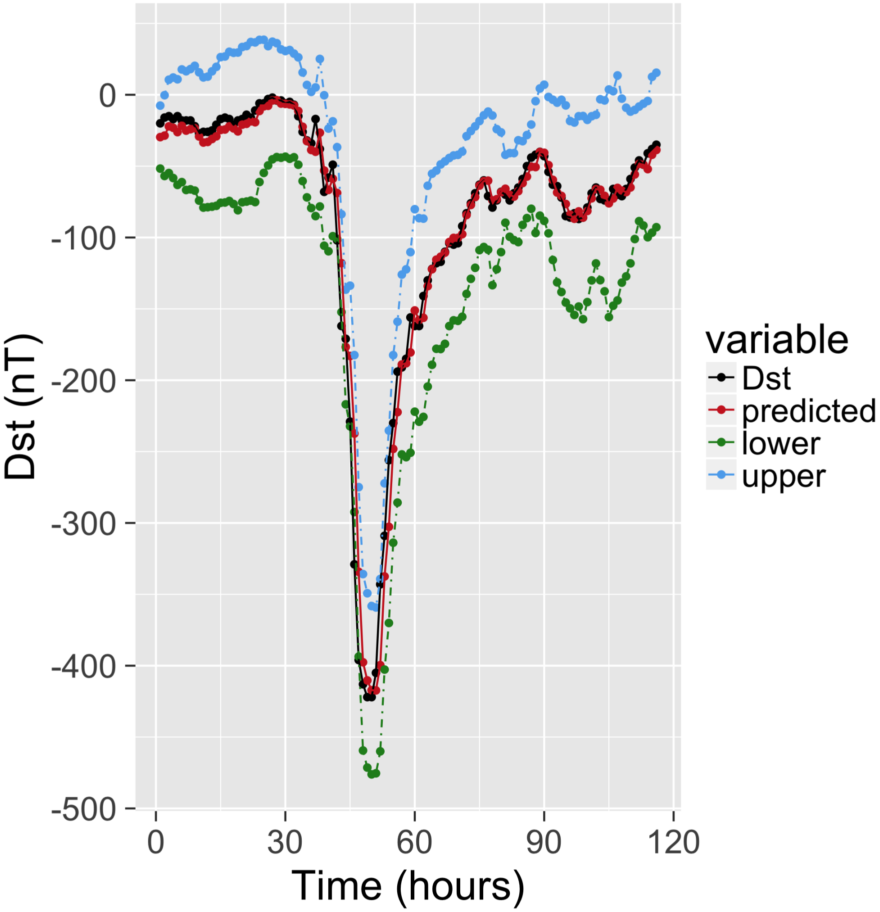


Figure 4.7: OSA Predictions with ±σ error bars for event: 2003/11/20 to 2003/11/22

Table 4.5: Storm events used to evaluate GP-AR and GP-ARX models

Event Id	**Start Date**	**Start Hour**	**End Date**	**End Hour**	**Storm Peak**
1	1998/02/17	12:00	1998/02/18	10:00	−100 nT
2	1998/03/10	11:00	1998/03/11	18:00	−116 nT
3	1998/05/04	02:00	1998/05/05	02:00	−205 nT
4	1998/08/26	10:00	1998/08/29	07:00	−155 nT
5	1998/09/25	01:00	1998/09/26	00:00	−207 nT
6	1998/10/19	05:00	1998/10/20	08:00	−112 nT
7	1998/11/09	03:00	1998/11/10	16:00	−142 nT
8	1998/11/13	00:00	1998/11/15	04:00	−131 nT
9	1999/01/13	16:00	1999/01/14	20:00	−112 nT
10	1999/02/18	03:00	1999/02/19	21:00	−123 nT
11	1999/09/22	20:00	1999/09/23	23:00	−173 nT
12	1999/10/22	00:00	1999/10/23	14:00	−237 nT
13	2000/02/12	05:00	2000/02/13	15:00	−133 nT
14	2000/04/06	17:00	2000/04/08	09:00	−288 nT
15	2000/05/24	01:00	2000/05/25	20:00	−147 nT
16	2000/08/10	20:00	2000/08/11	18:00	−230 nT
17	2000/08/12	02:00	2000/08/13	17:00	−235 nT
18	2000/10/13	02:00	2000/10/14	23:00	−107 nT
19	2000/10/28	20:00	2000/10/29	20:00	−127 nT
20	2000/11/06	13:00	2000/11/07	18:00	−159 nT
21	2000/11/28	18:00	2000/11/29	23:00	−119 nT
22	2001/03/19	15:00	2001/03/21	23:00	−149 nT
23	2001/03/31	04:00	2001/04/01	21:00	−387 nT
24	2001/04/11	16:00	2001/04/13	07:00	−271 nT
25	2001/04/18	01:00	2001/04/18	13:00	−114 nT
26	2001/04/22	02:00	2001/04/23	15:00	−102 nT
27	2001/08/17	16:00	2001/08/18	16:00	−105 nT
28	2001/09/30	23:00	2001/10/02	00:00	−148 nT
29	2001/10/21	17:00	2001/10/24	11:00	−187 nT
30	2001/10/28	03:00	2001/10/29	22:00	−157 nT
31	2002/03/23	14:00	2002/03/25	05:00	−100 nT
32	2002/04/17	11:00	2002/04/19	02:00	−127 nT
33	2002/04/19	09:00	2002/04/21	06:00	−149 nT
34	2002/05/11	10:00	2002/05/12	16:00	−110 nT
35	2002/05/23	12:00	2002/05/24	23:00	−109 nT
36	2002/08/01	23:00	2002/08/02	09:00	−102 nT
37	2002/09/04	01:00	2002/09/05	00:00	−109 nT
38	2002/09/07	14:00	2002/09/08	20:00	−181 nT
39	2002/10/01	06:00	2002/10/03	08:00	−176 nT
40	2002/10/03	10:00	2002/10/04	18:00	−146 nT
41	2002/11/20	16:00	2002/11/22	06:00	−128 nT
42	2003/05/29	20:00	2003/05/30	10:00	−144 nT
43	2003/06/17	19:00	2003/06/19	03:00	−141 nT
44	2003/07/11	15:00	2003/07/12	16:00	−105 nT
45	2003/08/17	18:00	2003/08/19	11:00	−148 nT
46	2003/11/20	12:00	2003/11/22	00:00	−422 nT
47	2004/01/22	03:00	2004/01/24	00:00	−149 nT
48	2004/02/11	10:00	2004/02/12	00:00	−105 nT
49	2004/04/03	14:00	2004/04/04	08:00	−112 nT
50	2004/07/22	20:00	2004/07/23	20:00	−101 nT
51	2004/07/24	21:00	2004/07/26	17:00	−148 nT
52	2004/07/26	22:00	2004/07/30	05:00	−197 nT
53	2004/08/30	05:00	2004/08/31	21:00	−126 nT
54	2004/11/07	21:00	2004/11/08	21:00	−373 nT
55	2004/11/09	11:00	2004/11/11	09:00	−289 nT
56	2004/11/11	22:00	2004/11/13	13:00	−109 nT
57	2005/01/21	18:00	2005/01/23	05:00	−105 nT
58	2005/05/07	20:00	2005/05/09	10:00	−127 nT
59	2005/05/29	22:00	2005/05/31	08:00	−138 nT
60	2005/06/12	17:00	2005/06/13	19:00	−106 nT
61	2005/08/31	12:00	2005/09/01	12:00	−131 nT
62	2006/04/13	20:00	2006/04/14	23:00	−111 nT
63	2006/12/14	21:00	2006/12/16	03:00	−147 nT

Chapter 5

Multiple Step Ahead Forecasts of the Disturbance Storm Time Index: The GPNN Model

We present a method that combines a *long short-term memory* (LSTM) network with a Gaussian process (GP) model to provide up to six-hour-ahead probabilistic forecasts of the Dst geomagnetic index. The proposed approach brings together the sequence modeling capabilities of a recurrent neural network with the error bars and confidence bounds provided by a GP. Our model is trained using the hourly OMNI and Global Positioning System (GPS) databases - both of which are publicly available. We first develop a LSTM network to get a single-point prediction of Dst. This model yields great accuracy in forecasting the Dst index from one to six hours ahead, with a correlation coefficient always higher than 0.873 and a root mean square error lower than 9.86 nT. However, even if its averaged metrics show excellent performance, it remains poor in predicting superstorms (Dst < -250 nT) six hours in advance. Next, we combine the LSTM model obtained from the first step with a GP model to yield a hybrid probabilistic predictor and evaluate it using the receiver operating characteristic curve and the reliability diagram. We conclude that this hybrid methodology provides improvements in the forecasting of Dst from one to five hours ahead.

This chapter is based on the following:

Article:
M. A. Gruet, M. Chandorkar, A. Sicard, and E. Camporeale. Multiple-hour-ahead forecast of the dst index using a combination of long short-term memory neural network and gaussian process. *Space Weather*, 16(11):1882–1896, 2018. doi: 10.1029/2018SW001898. URL `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2018SW001898`
Research work led by M. A. Gruet; M. Chandorkar provided guidance for the Gaussian process component as well as model tuning and evaluation experiments. This work has also been presented in Gruet [2018, ch. 5].

5.1 Introduction

It is widely accepted that solar wind magnetosphere coupling plays a key role in determining the Earth's geomagnetic state. Under appropriate conditions, this coupling can lead to injection of energetic particles into the Earth's auroral and equatorial plasma currents, leading to geomagnetic storms. The solar wind conditions that are effective for creating geomagnetic storms are sustained periods of high-speed solar wind and a southward directed solar wind magnetic field [Burton et al., 1975]. When Akasofu [1981] studied the coupling function between the solar wind and geomagnetic disturbances, they observed that during these extreme events, the key process is magnetic reconnection [Priest and Forbes, 2007]. Magnetic reconnection produces an enhancement of particle flux which creates a depression of the horizontal component of the Earth's magnetic field and an intensification of the westward ring current circulating the Earth [Gonzalez et al., 1994]. When there is a geomagnetic storm, the energy content of the ring current increases. This increase is inversely proportional to the strength of the surface magnetic field at low latitudes. To assess the severity of geomagnetic storms, the Disturbance Storm Time index (Dst) is often used.

The Dst index [Sugiura, 1964] is based on measurements from four low latitude stations and represents the axis-symmetric magnetic signature of magnetosphere currents (such as the ring current, the tail currents, and the Chapman-Ferraro current). It is computed using 1-hour average values of the horizontal component of the Earth's magnetic field and is expressed in nano Tesla (nT). In the case of a typical magnetic storm, three phases are observed according to Dst variations. First, there is a sudden drop corresponding to the storm commencement. Second, the value of Dst stays in its excited state as the ring current intensifies (the main phase). Finally, once the z-component of the interplanetary magnetic field (IMF) turns northward, the ring current begins to recover and rises back to its quiet level (recovery phase).

Geomagnetic indices like Dst, Kp, AE and others are used in space weather to describe effects of the solar wind on the Earth's magnetosphere. It has been long observed that strong geomagnetic storms disrupt man-made systems, they can impact satellites, disrupt navigation systems and create harmful geomagnetic induced currents in power grids and pipelines. One of the important research problems in space weather is to predict geomagnetic disturbances in order to protect technological infrastructure [Singh et al., 2010]. The aim of this study is to propose an accurate and reliable probabilistic model to predict Dst from one hour to six hours ahead.

The Dst prediction problem has been extensively researched. Burton et al. [1975] developed a model that expressed the time evolution of Dst as an Ordinary Differential Equation (ODE). This method takes into account the particle injection from the plasma sheet into the magnetosphere and expresses it based on the solar wind speed, density and on the north-south magnetic component of the IMF (B_z). Iyemori et al. [1979] used a linear filtering prediction method to connect Dst with the southward component of the interplanetary magnetic field. The linear assumption, however, has limitations since the solar wind and the magnetosphere form a coupled non-linear system.

To model nonlinear behavior of magnetospheric response, various models have been proposed. A popular approach used to model nonlinear systems is based on artificial neural networks (ANN) [Haykin, 1994]. One of the earliest models of Dst prediction based on ANNs is due to Lundstedt and Wintoft [1994]. They developed a feedforward neural network to predict Dst one hour ahead, using B_z, the density, and the velocity of the solar wind. This model was able to model the initial and main phases of geomagnetic storms well, but the recovery phase was not modeled accurately. Gleisner et al. [1996] developed a time delay neural network [Waibel et al., 1989] to predict Dst one hour ahead using the proton density, solar wind velocity, and Bz. This approach managed to improve the prediction of storm recovery phases, showing the benefits of using the time history of solar wind inputs. Wu and Lundstedt [1997] used an Elman recurrent network [Elman, 1990] to provide forecast of the Dst index from one hour to six hours ahead. Subsequently, Lundstedt et al. [2002] used the Elman network architecture to provide an operational forecast of the Dst index one hour ahead with improved performance. Wing et al. [2005] used a recurrent network, to provide an operational forecast of the Kp index. The success of these operational models demonstrate that recurrent networks are quite useful in the empirical modelling of magnetospheric response to solar wind drivers.

Another approach at the intersection between physical models and neural networks is provided by Bala and Reiff [2012]. Their approach is based on ANNs and uses the so called Boyle index which represents the steady state polar cap potential as an input. The Boyle index is a combination of the velocity of the solar wind, the magnitude of the IMF, and the IMF clock angle; it can be used to predict Kp, Dst, and AE from one to six hours ahead with good results. Lazzús et al. [2017] use Particle Swarm Optimization (PSO) [Eberhart and Kennedy, 1995], instead of the backpropagation algorithm [Rummelhart et al., 1986], to learn the ANN connection weights. Results obtained in this study show that PSO can provide benefits for training Dst prediction models.

Chandorkar et al. [2017] point out that although several techniques have been used to predict Dst, these models do not focus on providing probabilistic predictions. Their model is based on *Gaussian processes* (GP) to construct autoregressive models to predict Dst one hour ahead, based on a time history of Dst, solar wind velocity, and the IMF B_z. They demonstrated that it is possible to generate an accurate predictive distribution of the forecast instead of a single point prediction. This is important in the space weather domain where operators require error bars on predictions. However, the mean value of the forecast does not yield a performance as accurate as the one provided by ANN.

All the models discussed above are based either on solar wind parameters or past values of Dst. One of the most striking features of the Dst index is the link between its variation and the impact it has on GPS satellites. It is widely known that when there is a geomagnetic storm, the quality of the GPS signal is disturbed [Astafyeva et al., 2014]. The magnetic field measured onboard GPS satellites might be key information when an important storm occurs [Morley et al., 2017]. Recently, GPS data has been publicly released under the terms of the *executive order for coordinating efforts to prepare the nation for space weather events* [The White House, 2016].

In this work, we propose a technique which combines the predictive capabilities of an ANN with the advantages of the probabilistic forecast provided by a GP. We use the well known long short-term memory network (LSTM) [Hochreiter and Schmidhuber, 1997] to provide a single point prediction of the Dst geomagnetic index from one to six hours ahead. The prediction produced by the LSTM is used as the mean function of a GP, to obtain a probabilistic forecast. We call this hybrid model the *Gaussian process neural network* (GPNN). Input parameters of the GPNN are solar wind parameters (density, velocity, IMF$|B|$ and Bz), six hour time history of Dst, and the magnetic field measured onboard GPS satellites. To the best of our knowledge, the LSTM architecture has never been used in space weather applications before.

The remainder of this chapter is organised into sections as follows. section 5.2 presents the data used in this study, section 5.3 describes the computational method and how the LSTM and its combination with the GP are developed and optimised. Section 5.4 presents the performance of the LSTM model when producing point predictions for the Dst one to six hours ahead and the evaluation of the probabilistic forecasts provided by the GPNN model.

5.2 Data

The solar wind parameters and the geomagnetic Dst index are taken from the OMNI data set[1] maintained by the National Space Science Data Center (NSSDC) of National Aeronautics and Space Administration (NASA).

We also consider GPS data made available by the National Oceanic and Atmospheric Administration (NOAA). The Combined X-ray dosimeter (CXD) team at the Los Alamos National Laboratory[2] provided this data set. In this study, we used measurements recorded by the GPS satellite ns41, which has the widest temporal coverage [Morley et al., 2017].

Figure 5.1 shows the temporal coverage of the database used in this study compared with previous studies. The temporal coverage of our study is represented by the green line. As GPS ns41 data starts at $00:00$ 14 January 2001, we consider a set of $134,398$ hourly data points consisting of solar wind parameters, geomagnetic Dst index, and GPS data between this starting date and $23:00$ 31 December 2016. This includes 49 storm events, which are listed in table 5.9. Some of these events in table 5.9 also overlap with events included in Ji et al. [2012] and Chandorkar et al. [2017].

---

[1] `https://omniweb.gsfc.nasa.gov/ow.html`

[2] `https://www.ngdc.noaa.gov/stp/space-weather/satellite-data/satellite-systems/gps/`

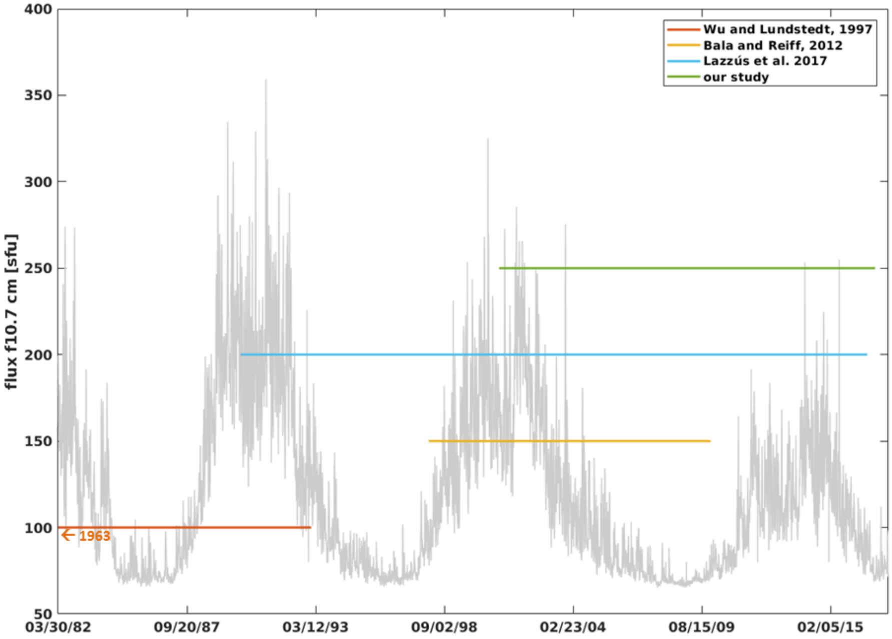


Figure 5.1: Temporal coverage of database used in this study and in previous studies. Wu and Lundstedt [1997] is in orange and their database starts in 1963, Bala and Reiff [2012] is in yellow, Lazzús et al. [2017] is in blue, and our study is in green. The f10.7 in grey represents the variation of solar activity.

Studies done in the past to predict the geomagnetic index Dst have shown that various solar wind parameters are of interest when improving performance of Dst models. In the present study, we focused on the use of the solar wind density n, velocity V_{sw}, the interplanetary magnetic field strength $||\mathbf{B}||$, and its north-south component B_z. Concerning parameters provided by the GPS satellite ns41, we use its magnetic field measurement, B_{gps}.

5.3 Methodology

5.3.1 Long Short-Term Memory Network

The long short-term memory network (LSTM) belongs to the family of *recurrent neural networks* (RNN). In an RNN, hidden layers are built to allow information persistence. They behave as a loop to allow information to be passed from one cell of the network to the next. When this loop is unrolled, the RNN can then be thought as multiple copies of the same network. The RNN architecture is particularly suited for time series forecasting applications.

Hochreiter [1991], Bengio et al. [1994] underlined a weakness of RNNs. They are supposed to connect past information to the present, but if the information needed is too far in the past, RNNs are unable to to retain it. This failure is due to the well known *vanishing gradient problem* occurring during the training of RNNs.

LSTM networks are designed to overcome the vanishing gradient problem by retaining information pertaining to long range temporal dependencies in time series data. LSTMs have a chain-like structure like RNNs, but the repeating module known as the LSTM cell has specific characteristics.

Figure 5.2 shows the computational graph of the LSTM cell. The two elements fundamental to this cell are the cell state and the gates. The cell state $\mathbf{c}_t$ in figure 5.2 is like a conveyor belt which is connected to various gates. Gates can add or remove information from the cell state depending on the information required by the LSTM unit. The important pieces of this architecture are: 1. the forget gate $\mathbf{f}_t$, 2. the input gate $\mathbf{i}_t$, 3. the cell state $\mathbf{c}_t$, and 4. the output gate $\mathbf{o}_t$

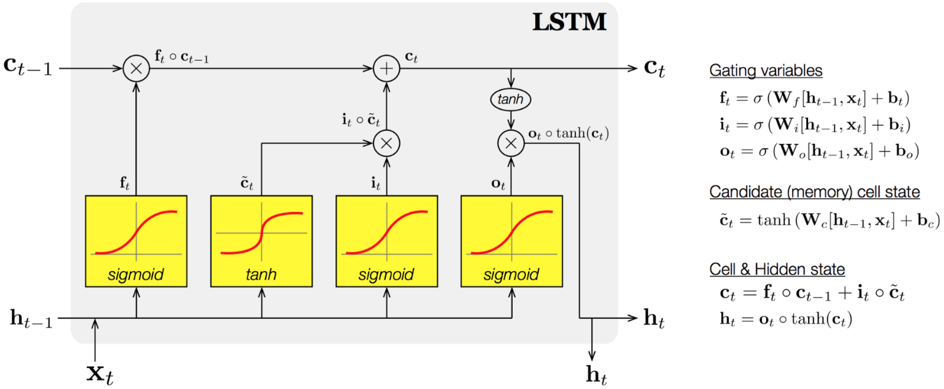


Figure 5.2: Schematic diagram of the LSTM cell. Reproduced from `https://github.com/llSourcell/LSTM_Networks/blob/master/LSTM%20Demo.ipynb`

Forget Gate & Input Gate

The forget gate is expressed as

$$\mathbf{f}_t = \sigma\left(\mathbf{W}_f \cdot [\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_f\right) \ , \tag{5.1}$$

where σ is the component wise sigmoid function and $\mathbf{W}_f$ and $\mathbf{b}_f$ are the weight matrix and bias vector of the gate respectively. This notation is kept for subsequent equations. The forget gate compares the information coming from the previous cell $\mathbf{h}_{t-1}$ and the incoming information $\mathbf{x}_t$, and outputs a number between zero and one. Zero is achieved if the information is entirely discarded, and one is given if it is altogether retained. Correspondingly, the input gate in equation (5.2) decides what information is retained, depending on past hidden cell state $\mathbf{h}_{t-1}$ and exogenous inputs $\mathbf{x}_t$. $\mathbf{W}_i$ and $\mathbf{b}_i$ are the weight matrix and bias vector of the input gate respectively.

$$\mathbf{i}_t = \sigma\left(\mathbf{W}_i \cdot [\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_i\right) \tag{5.2}$$

Cell State

The inputs $\mathbf{x}_t$ and cell hidden state $\mathbf{h}_{t-1}$ are then passed through a hyperbolic tangent transformation to create a vector of candidate values $\tilde{\mathbf{c}}_t$ as per equation (5.3). This candidate state $\tilde{\mathbf{c}}_t$ is used to create the updated cell state $\mathbf{c}_t$, $\mathbf{W}_c$ and $\mathbf{b}_c$ being the weight and bias of this layer respectively.

$$\tilde{\mathbf{c}}_t = \tanh\left(\mathbf{W}_c \cdot [\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_c\right) \tag{5.3}$$

The new cell state $\mathbf{c}_t$ is computed as a weighted sum of the old cell state $\mathbf{c}_{t-1}$ and the candidate cell state $\tilde{\mathbf{c}}_t$. This is expressed as

$$\mathbf{c}_t = \mathbf{f}_t \circ \mathbf{c}_{t-1} + \mathbf{i}_t \circ \tilde{\mathbf{c}}_t \ , \tag{5.4}$$

where $\circ$ denotes the element wise (Hadamard) product. The old cell state is appropriately 'forgotten' using the output of $\mathbf{f}_t$, and information from the candidate state is allowed to enter $\mathbf{c}_t$ using $\mathbf{i}_t$.

Output Gate & Prediction

The output of the LSTM cell is calculated using equation (5.5). First, the sigmoid transformation helps to define the output $\mathbf{o}_t$. Second, the hidden state $\mathbf{h}_t$ is computed by multiplying $\mathbf{o}_t$ with a hyperbolic tangent transformation of the cell state $\mathbf{c}_t$. The value of $\mathbf{h}_t$ (seen in figure 5.2) becomes the final prediction computed LSTM cell for input $\mathbf{x}_t$.

$$\mathbf{o}_t = \sigma\left(\mathbf{W}_o \cdot [\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_o\right) \tag{5.5}$$

$$\mathbf{h}_t = \mathbf{o}_t \circ \tanh(\mathbf{c}_t) \tag{5.6}$$

5.3.2 The LSTM Dst Model

We trained six LSTM models, one for each hour in the forward time window. Thus for every input, we obtained a vector of outputs $[\hat{\mathrm{Dst}}\,(t+p)], p \in \{1, \cdots, 6\}$. For inputs $\mathbf{x}_t$, we use solar wind parameters (n, V_{sw}, $|\mathbf{B}|$, B_z) and the GPS data (B_{gps}) described in section 5.2. We also use the time history of Dst, from one to six hours back.

$$\begin{aligned}\mathbf{x}_t = \big(&n\,(t)\,, V_{\mathrm{sw}}\,(t)\,, |\mathbf{B}\,(t)\,|, B_z\,(t)\,, B_{\mathrm{gps}}\,(t)\,,\\ &\mathrm{Dst}\,(t-1)\,, \mathrm{Dst}\,(t-2)\,, \ldots, \mathrm{Dst}\,(t-6)\big)\end{aligned} \tag{5.7}$$

Architecture & Implementation

LSTM cells can be *stacked* such that the output of one cell provides the input for the next. One may construct LSTM stacks of increasing size depending on the task at hand.

In equations (5.8) and (5.9), we describe the Dst prediction architecture. The p hour ahead prediction $\hat{\mathrm{D}}\mathrm{st}\,(t+p)$ is the output of the function $m^{p}_{\mathrm{NN}}(\cdot)$, which is obtained by successive operation of d LSTM cells. Mathematically this is equivalent to function composition $h^{p}_{d}(h^{p}_{d-1}(\cdots(h^{p}_{1}(\cdot)\cdots))$.

$$\hat{\mathrm{D}}\mathrm{st}\,(t+p) = m^{p}_{\mathrm{NN}}(\mathbf{x}_t) \tag{5.8}$$

$$m^{p}_{\mathrm{NN}}(\mathbf{x}_t) = h^{p}_{d}(h^{p}_{d-1}(\cdots(h^{p}_{1}(\mathbf{x}_t)\cdots)) \tag{5.9}$$

To find the LSTM structure, which is most suitable for predicting geomagnetic activity, we built networks consisting of varying number of cells. After training and validating each candidate architecture, the best performing LSTM network was chosen as the mean function for the Gaussian process component outlined in section 5.3.4.

The LSTM component of the GPNN model was implemented using the Lasagne library in Python [Dieleman et al., 2015, Theano Development Team, 2016].

5.3.3 LSTM Model Training

The LSTM architecture can be trained with an iterative backpropagation based optimization algorithm. Unlike the case of RNN network training, the vanishing gradient problem does not persist. A number of stochastic gradient based variants are used for training neural networks, such as, but not limited to, *Levenburg - Maquardt* [Marquardt, 1963], *Nesterov Accelerated Gradient* (NAG) [Nesterov, 1983], adaptive learning rates for each network weight [Silva and Almeida, 1990], adaptive gradient based methods such as AdaGrad [Duchi et al., 2011], and adaptive learning rate methods like RMSprop [Tieleman and Hinton, 2012].

In this work, we use the RMSProp algorithm for training the LSTM component of the GPNN model. Below we give a brief explanation of its functioning.

The RMSProp Algorithm

For the purposes of notational simplicity, the weights and biases of the LSTM architecture are concatenated into a single vector θ, where θ_i denotes an individual scalar element of θ and θ_i^k the value of θ_i at iteration k. The quantity $J(\theta)$ is the objective function which defines how well the outputs of the LSTM model fit the observed data. The training process of the LSTM consists of computing an approximate minimum of $J(\theta)$. We define $g_{k,i}$ in equation (5.10) as the gradient of the objective function with respect to the parameter θ_i at iteration k.

$$g_{k,i} = \left.\frac{\partial J(\theta)}{\partial \theta_i}\right|_{\theta=\theta^k} \tag{5.10}$$

The aim of the training process is to successively decrease the value of $J(\theta)$ over a number of training iterations. This is achieved as follows.

First, the running average of the second order moment of the gradients $\mathrm{E}\left(g^2\right)$ is computed at each iteration k. Then, the updated parameters θ_i^{k+1} are calculated using the gradient $g_{k,i}$ and a damped learning rate of $\frac{\eta}{\sqrt{\mathrm{E}[g^2]_{k,i}+\epsilon}}$ (ϵ is a small number added to $\sqrt{\mathrm{E}\left[g^2\right]_{k,i}}$ to prevent numerical overflow). The update of parameters using RMSprop can then be computed using equation (5.11).

$$\mathrm{E}\left[g^2\right]_{k+1,i} = 0.9\mathrm{E}\left[g^2\right]_{k,i} + 0.1 g_{k,i}^2 \tag{5.11}$$

$$\theta_i^{k+1} = \theta_i^k - \frac{\eta}{\sqrt{\mathrm{E}\left[g^2\right]_{k,i} + \epsilon}} g_{k,i} \tag{5.12}$$

5.3.4 Gaussian Processes

Gaussian processes are a family of models that provide a principled probabilistic framework for forecasting. GP models output a predictive distribution instead of a point forecast. Starting from a *prior distribution*, GP models construct a *posterior predictive distribution*. The appeal of using GP models is that, their practical implementation is straightforward boiling down to a simple analytical expression that requires no more than linear algebra, although their theoretical formulation is general. In this study, the

GP component is implemented using the GPML Matlab software package [Rasmussen and Nickisch, 2010].

We will not give a complete review of the GP methodology here. The interested reader can refer to section 4.2 for a detailed review of Gaussian process models and their mathematical formulation.

A GP is completely specified by its mean function $m\left(\mathbf{x}\right)$ and its covariance function or *kernel* $K(\mathbf{x}, \mathbf{x}')$ in equation (4.8).

Kernel Function

Kernels determine how each point $\mathbf{x}_t$ influences the values that the GP will have on another point $\mathbf{x}_s$. The main idea is that if $K(\mathbf{x}_t, \mathbf{x}_s) \gg 0$, we expect the output from the GP model at these points to be statistically correlated.

$$K_{\mathrm{NN}}\left(\mathbf{x}, \mathbf{x}'\right) = \frac{2}{\pi}\sin^{-1}\left(\frac{2\mathbf{x}\cdot\mathbf{x}'}{\sqrt{(1+2\mathbf{x}\cdot\mathbf{x})}\sqrt{(1+2\mathbf{x}'\cdot\mathbf{x}')}}\right) \tag{5.13}$$

Kernels often introduce assumptions about the continuity of the functions that will be represented by the resulting GP regression model. Rasmussen and Williams [2005, ch. 4] gives an introduction to common kernels used in machine learning applications and the type of continuity assumptions of their resulting GP models. In this study, we use the neural network kernel [Williams, 1998] described in equation (5.13).

Mean Function

The mean function $m\left(\mathbf{x}\right)$ defines an a priori mean value of the GP predictive distribution for some input $\mathbf{x}$. It is quite common to set $m(\mathbf{x}) = 0$ without loss of generality for situations where it is difficult to define an a priori expression for the predictive mean. Rasmussen and Williams [2005, ch. 2,sec. 2.7] states that using mean functions is a good way to incorporate domain knowledge or to augment the model's capabilities in capturing complex behavior. This outlines the structure of the hybrid GPNN model consisting of an LSTM and GP components.

5.3.5 The GPNN Model

$$\mathrm{Dst}\,(t+p) = f_p\,(\mathbf{x}_t) + \epsilon \tag{5.14}$$
$$\epsilon \sim N\left(0, \sigma^2\right) \tag{5.15}$$
$$f_p\,(\mathbf{x}_t) \sim \mathcal{GP}\left(m^p_{\mathrm{NN}}(\mathbf{x}_t), K_{NN}(\mathbf{x}_t, \mathbf{x}_s)\right) \tag{5.16}$$

In equations (5.14) to (5.16), we give a formal specification of the proposed GPNN model in terms of the notations introduced in section 4.2. The ground truth $\mathrm{Dst}\,(t+p)$ is assumed to be the result of a Gaussian Process f_p corrupted by Gaussian measurement noise ϵ, with $p = 1, 2, 3, 4, 5, 6$ being the forecast horizon going from one to six hours ahead. The mean function of the Gaussian process f_p is set to $m^p_{\mathrm{NN}}\,(\mathbf{x})$ described in section 5.3.2 and the covariance is set to the neural network covariance function from equation (5.13).

5.4 Experiments

5.4.1 LSTM Model Evaluation

For the purposes of training and evaluation, the data set is divided as follows: 70% for training, 20% for testing and 10% for validation. The LSTM model is evaluated using the root mean square error (RMSE) and the Pearson correlation coefficient (CC) defined by equations (5.17) and (5.18) respectively. Based on the performance calculated on the validation data, an LSTM network of depth $d = 20$ cells was chosen as the mean function of the GPNN model.

$$\mathrm{RMSE} = \sqrt{\sum_{t=1}^{n} \left(\mathrm{Dst}\,(t) - \hat{\mathrm{D}}\mathrm{st}\,(t)\right)^2 / n} \tag{5.17}$$

$$\mathrm{CC} = \frac{\mathrm{Cov}(\mathrm{Dst}, \hat{\mathrm{D}}\mathrm{st})}{\sqrt{\mathrm{Var}(\mathrm{Dst})\,\mathrm{Var}(\hat{\mathrm{D}}\mathrm{st})}} \tag{5.18}$$

In order to give context for the performance of our LSTM model, we compare the following models for Dst predictions for the time horizon $t+1 \cdots t+6$, summarised in table 5.1.

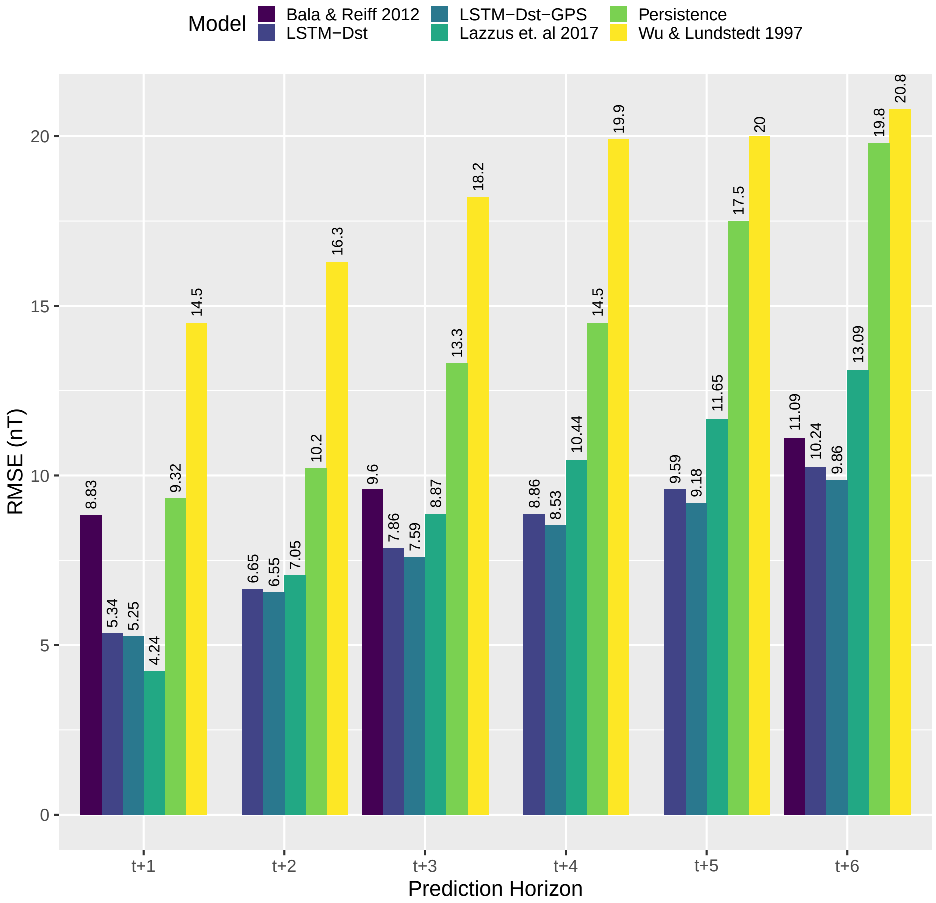


Figure 5.3: RMSE comparison of Dst forecast models.

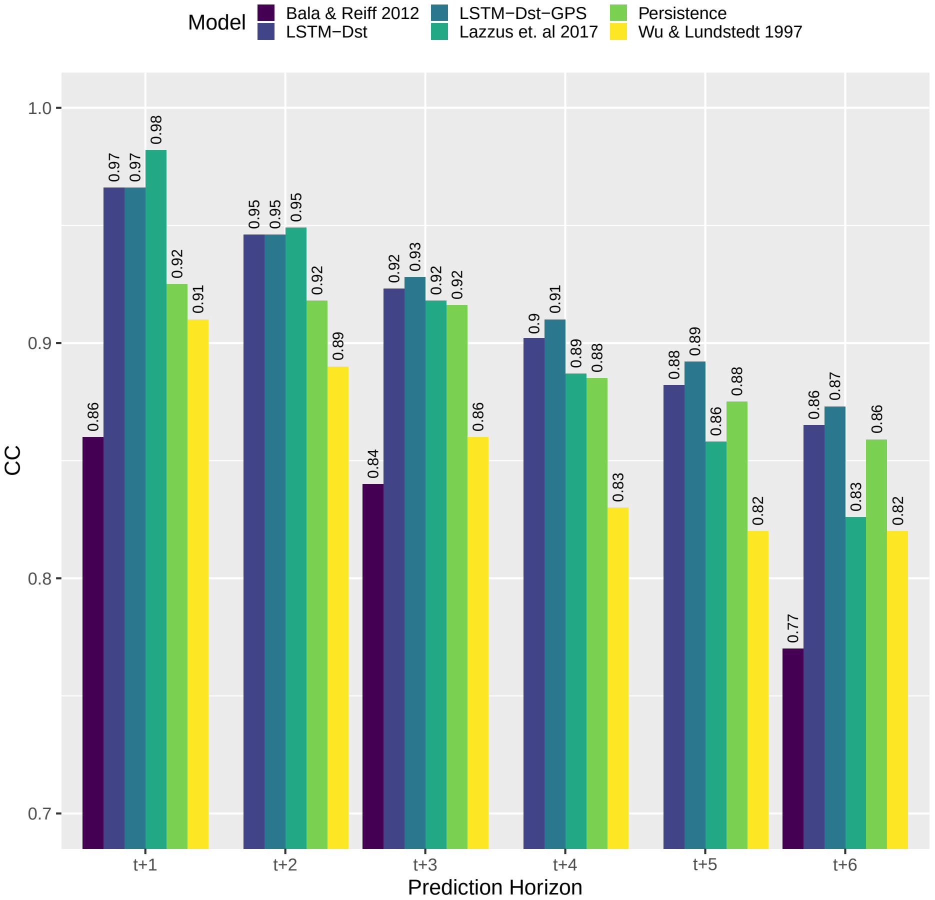


Figure 5.4: CC comparison of Dst forecast models.

Table 5.1: Summary of Dst forecasting models compared in this study

Model	**Description**
Wu and Lundstedt [1997]	An Elman RNN model having explicitly computed solar wind magnetospheric coupling functions as inputs.
Bala and Reiff [2012]	A feedforward neural network model using the *Boyle* function as input.
Lazzús et al. [2017]	A feedforward neural network model using only the time history of Dst. Trained using particle swarm optimization.
Persistence model	$\hat{\text{Dst}}(t+p) = \text{Dst}(t)$, also used for benchmarking in section 4.4.
LSTM-Dst	Our proposed LSTM model. Inputs are solar wind parameters (section 5.2) and time history of Dst.
LSTM-Dst-GPS	The LSTM model but with GPS magnetic field data B_{gps} also included in the inputs.

Figure 5.4 present a comparison of CC and RMSE performance of the Dst prediction models listed in table 5.1.

Our models LSTM-Dst & LSTM-Dst-GPS provide performance which is competitive to that obtained by Lazzús et al. [2017] for one to three hours ahead. When the forecast horizon goes from four to six hours ahead, LSTM-Dst & LSTM-Dst-GPS provide better global performance. As an example, when considering a six-hour-ahead forecast, LSTM-Dst-GPS provides a CC of 0.873 and a RMSE of 9.86 nT, while Lazzús et al. [2017] obtained a CC of 0.826 and a RMSE of 13.09 nT. The model presented in Lazzús et al. [2017] uses only time history of Dst. This emphasises the performance benefits of using exogenous data when predicting geomagnetic activity with a time horizon greater than one hour.

Bala and Reiff [2012] used the Boyle index as an input function and obtained competitive RMSE performance to LSTM for six hour ahead predictions. Their model presents a CC of 0.77 and a RMSE of 11.09 nT. The Elman RNN proposed in Wu and Lundstedt [1997] is an important benchmark for the GPNN model because the LSTM architecture is a more sophisticated form of the RNN structure. Wu and Lundstedt [1997] pro-

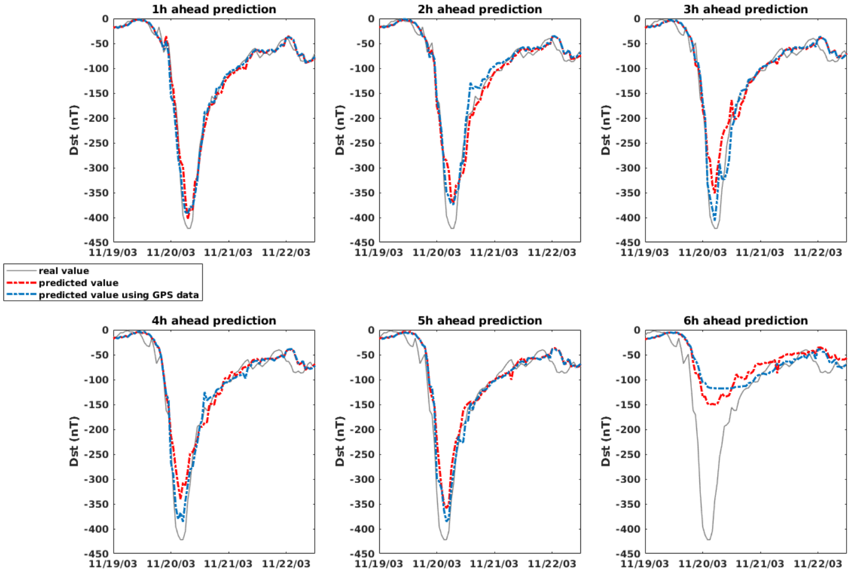


Figure 5.5: Dst predictions made by the LSTM-Dst and LSTM-Dst-GPS models for the 2003 Halloween storm.

vided for a six ahead forecast, a CC of 0.82 and a RMSE of 20.8 nT. The LSTM architecture thus offers performance benefits over standard RNNs when making geomagnetic activity.

We observed that using GPS data generally results in an improved forecasts of important geomagnetic storm events. Figure 5.5 presents forecasts produced the LSTM-Dst in blue and LSTM-Dst-GPS in red for the 2003 Halloween storm event (maximum strength of −422 nT). Predictions for one to two hours ahead are very similar, but when we consider the forecasts made for three hours ahead, LSTM-Dst predicts a storm peak of −348 nT while LSTM-Dst-GPS provides a prediction of −405 nT.

For a four-hour-ahead forecast, LSTM-Dst predicts a storm peak of −335 nT while LSTM-Dst-GPS predicts −380 nT.

5.4.2 GPNN Evaluation

Since GP models output a predictive distribution, metrics like RMSE and CC, which are defined for single point predictions, are not adequate for evaluating probabilistic forecasts.

Table 5.2: Storm Classification

Level of Activity	**Storm Classification**
$\text{Dst} > -50\,\text{nT}$	Moderate
$-250\,\text{nT} \leq \text{Dst} \leq -50\,\text{nT}$	Intense
$\text{Dst} < -250\,\text{nT}$	Super Storm

The GPNN model provides an operator a probabilistic forecast, which can be important in a decision making scenario. For example, a satellite operator may choose to turn off a sensitive components or equipment before an intense storm, when the forecast probability of an intense event exceeds a predetermined trigger threshold.

Storm activity is often classified by thresholding Dst values. In line with the most common classification schemes, we distinguished three levels of geomagnetic activity summarised in table 5.2. It is ideal to use metrics that will be able to evaluate how well the GPNN model manages to correctly classify geomagnetic storms one to six hours prior. To do so, we used receiver operating characteristic (ROC) curves and reliability diagrams.

Receiver Operating Characteristic Curve

The ROC curve is based on a contingency table which maps how often a model correctly classifies outcomes when working in a binary classification setting i.e. occurrence or non occurrence of a predefined event, denoted by labels 1 and 0 respectively. From table 5.2, we constructed three binary classification problems for each storm class.

In the case of the GPNN model, we can analytically compute the probability that Dst lies within a particular storm class or outside it (e.g. $\mathbb{P}[-250\,\mathrm{nT} \leq \mathrm{Dst} \leq -50\,\mathrm{nT}]$ is the probability of occurrence of an intense storm). For any storm class, by choosing a threshold probability, we can assign a classification of 1 (belonging to that storm class) if the probability predicted by the GPNN model is greater than the chosen threshold, and 0 if otherwise.

The ROC contingency table is then constructed by tabulating the relationship between the *true positive ratio* (TPR) and *false positive ratio* for varying threshold probabilities. A *true positive* is when an occurrence of the event in question is correctly classified by the model as such. A *false positive* is when a non-occurrence is classified as an occurrence. The TPR is then the ratio of true positives to the total number of event occurrences, while the FPR is the ratio of false positives to the number of non occurrences. For perfect classification, FPR $= 0$ and TPR $= 1$. Thus, the value of the threshold that produces the point closest to these values is optimal.

In tables 5.3 to 5.8, we present ROC curves obtained from one to six hour ahead forecasts, organised by storm category. The ROC is usually shown graphically, but numerical values are more relevant for the reader to analyse variations depending on the chosen threshold probability. The optimal TPR and FPR values are in bold, obtained by minimising the Euclidean distance from FPR $= 0$ and TPR $= 1$.

For predictions done from one to five hours ahead, TPR values are always greater than 0.719 for thresholds from 10% to 40%, and then decrease with increasing thresholds. If we focus on the six-hour-ahead forecast, the best TPR is 0.5 for a 10% threshold. It means that the more there is an increasing probability for a superstorm to occur, the less the model is able to forecast it without misjudgments six hours in advance. However, for intense storms ($-250\,\mathrm{nT} < \mathrm{Dst} < -50\,\mathrm{nT}$), the GPNN has a TPR higher than 0.670 for thresholds between 10% and 80%, and for moderate storms, this model has a TPR higher than 0.649 for all thresholds, for predictions going from one to six hours ahead.

Reliability diagram

The ROC tables discussed in the previous section give information about the ability of the forecasting system to detect the occurrence of geomagnetic storm events based on a chosen decision threshold, in terms of false and true positives. Reliability diagrams measure how closely the forecast probabilities of an event correspond to the actual frequency with which an event is observed. A perfectly reliable forecast is one in which an event predicted with probability p is observed, on average, with frequency p. The reliability diagram bins the forecasts into groups according to the predicted probability, which is shown on the horizontal axis. The frequency with which an event was observed to occur for each bin is then plotted on the vertical axis. If the reliability curve lies above/below than the perfect diagonal slope, the resulting forecasts are under/over confident (i.e. they yield smaller/higher probabilities for a specific outcome than observed).

Figure 5.6 presents reliability diagrams obtained from one to six-hour-ahead forecasts made by the GPNN. It shows that the one-hour-ahead forecast slightly underestimates the likelihood of storm events when their occurrence probability is 35%. For example, when there is 80% predicted chance of a storm, the real observed frequency of it is 90%. The GPNN provides reliable forecast for two-hour-ahead predictions, as the observed frequency of storm regarding the predicted probability lies almost perfectly on the diagonal. For predictions further than three hours ahead, GPNN tends to overestimate the probability of storms.

If we focus on the six-hour-ahead prediction, when the GPNN model provides a predicted probability of 90%, the real frequency is 65%. GPNN $t + 6$ model is thus overconfident.

Figure 5.7 presents predictions provided by the GPNN model for the 2003 Halloween storm. For predictions from one to five hours ahead, GPNN provides accurate predictions with plausible error bars. For example, for the five hours ahead forecast, the storm peak of $-422\,\mathrm{nT}$ is forecasted as $-391\,\mathrm{nT}$.

The main benefit of the GP component in GPNN can be appreciated when comparing figures 5.5 and 5.7, which show that the GPNN model

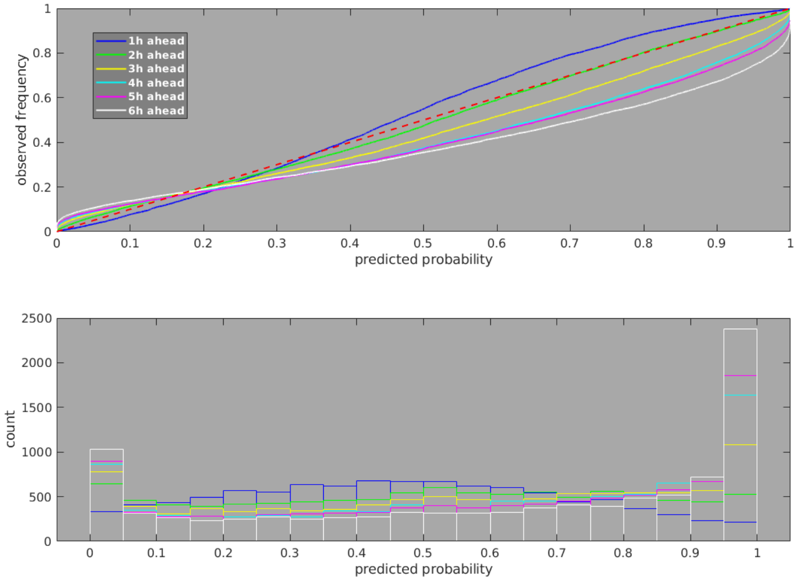


Figure 5.6: Reliability diagram for Dst forecast from one to six hours ahead. The diagonal is in red dot line.

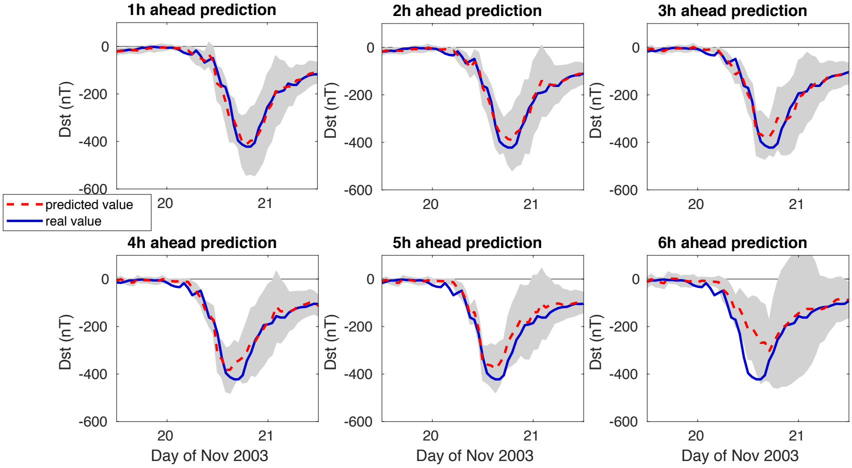


Figure 5.7: Dst predictions made by the GPNN for the 2003 Halloween storm. The most probable prediction is the dotted purple line. The ground truth Dst is the deep blue line. The grey shadow represents $\pm\sigma$ error bars on the prediction.

gave more accurate six-hour-ahead forecasts for the 2003 Halloween storm. The error bars calculated by the GPNN model include the storm peak for one to six-hour-ahead forecasts.

5.5 Conclusions

In this chapter, we presented a hybrid model for Dst forecasting, the GPNN, based on LSTM networks and Gaussian Processes.

First, we developed a LSTM network to provide Dst predictions one to six hours ahead. A separate LSTM was developed for each time step of the forecast horizon, then the global performance of LSTM was compared to previously proposed neural network based Dst forecast models. We observed that the LSTM provides competitive performance in comparison to the state of the art in Dst forecasting. When focusing on extreme events like the well known 2003 Halloween storm, we noted that even though our model's global performance metrics are excellent, the six-hour-ahead forecast fails to anticipate the storm peak.

Second, to obtain probabilistic forecasts, we developed a GP which encapsulates the LSTM model as the mean function. Thanks to this combination, we observed that the GPNN model managed to give accurate predictions for the 2003 Halloween storm for a one to five hour time horizon. For the six-hour-ahead prediction, the GPNN's predictive variance managed to encompass the storm peak.

To evaluate this probabilistic forecast, we used ROC curves and reliability diagrams. The ROC curves demonstrated that, for each forecast horizon, storm category and acceptance threshold, the FPR is low. The TPR values are excellent for moderate and intense storms, but for six-hour-ahead prediction of superstorms, misjudgment is possible when the acceptance threshold increased. For six hour ahead prediction, the optimal acceptance threshold is around 10%, which showed room for further improvement. The reliability diagram showed that the GPNN provides great performance for predictions from one to three ahead, but for four to six hours ahead, an overestimation of storm likelihoods is possible.

Table 5.3: ROC contingency table for the $t + 1$ GPNN model.

1-hr-ahead prediction						
	Super Storm		*Intense Storm*		*Moderate*	
Threshold	**TPR**	**FPR**	**TPR**	**FPR**	**TPR**	**FPR**
10%	0.969	2.70×10^{-3}	0.981	0.163	0.999	0.434
20%	0.969	1.11×10^{-3}	0.961	0.105	0.996	0.321
30%	**0.969**	**6.40×10^{-4}**	**0.927**	**0.0719**	0.991	0.240
40%	0.969	4.00×10^{-4}	0.895	0.049	0.984	0.185
50%	0.844	3.00×10^{-4}	0.855	0.0270	0.972	0.138
60%	0.812	2.78×10^{-4}	0.806	0.0161	0.951	0.102
70%	0.656	2.78×10^{-4}	0.753	$9.30.10^{-3}$	**0.929**	**0.0705**
80%	0.625	2.78×10^{-4}	0.670	$3.95.10^{-3}$	0.895	0.0371
90%	0.468	9.27×10^{-5}	0.554	$1.61.10^{-3}$	0.838	0.0178

Table 5.4: ROC contingency table for the $t + 2$ GPNN model.

2-hr-ahead prediction						
	Super Storm		*Intense Storm*		*Moderate*	
Threshold	**TPR**	**FPR**	**TPR**	**FPR**	**TPR**	**FPR**
10%	0.969	3.15×10^{-3}	0.963	0.199	0.999	0.388
20%	0.937	9.27×10^{-4}	0.934	0.142	0.984	0.273
30%	**0.937**	**3.71×10^{-4}**	**0.914**	**0.105**	0.973	0.211
40%	0.906	1.85×10^{-4}	0.891	0.0834	0.961	0.167
50%	0.781	1.85×10^{-4}	0.863	0.0565	0.943	0.134
60%	0.6875	9.27×10^{-5}	0.824	0.0390	**0.917**	**0.107**
70%	0.656	9.27×10^{-5}	0.783	0.0268	0.895	0.0845
80%	0.500	9.27×10^{-5}	0.720	0.0156	0.858	0.0646
90%	0.437	0	0.601	$5.6810^{-}3$	0.802	0.0363

Table 5.5: ROC contingency table for the $t + 3$ GPNN model.

3-hr-ahead prediction						
	Super Storm		*Intense Storm*		*Moderate*	
Threshold	**TPR**	**FPR**	**TPR**	**FPR**	**TPR**	**FPR**
10%	0.875	3.24×10^{-3}	0.958	0.254	0.984	0.373
20%	**0.843**	$\mathbf{9.27 \times 10^{-4}}$	0.939	0.186	0.971	0.278
30%	0.813	4.64×10^{-4}	0.912	0.139	0.955	0.228
40%	0.750	1.86×10^{-4}	**0.890**	**0.106**	0.940	0.182
50%	0.625	9.27×10^{-5}	0.880	0.0819	0.919	0.146
60%	0.593	0	0.809	0.0606	**0.893**	**0.1058**
70%	0.593	0	0.766	0.0451	0.826	0.0865
80%	0.437	0	0.714	0.0291	0.814	0.0594
90%	0.406	0	0.614	0.0164	0.747	0.0413

Table 5.6: ROC contingency table for the $t + 4$ GPNN model.

4-hr-ahead prediction						
	Super Storm		*Intense Storm*		*Moderate*	
Threshold	**TPR**	**FPR**	**TPR**	**FPR**	**TPR**	**FPR**
10%	0.906	3.24×10^{-3}	0.968	0.311	0.970	0.339
20%	**0.875**	$\mathbf{1.29 \times 10^{-3}}$	0.953	0.252	0.949	0.243
30%	0.813	7.42×10^{-4}	0.933	0.208	0.931	0.192
40%	0.813	6.49×10^{-4}	**0.916**	**0.169**	0.906	0.144
50%	0.781	9.27×10^{-5}	0.895	0.138	**0.874**	**0.104**
60%	0.687	9.27×10^{-5}	0.843	0.106	0.841	0.0803
70%	0.562	9.27×10^{-5}	0.795	0.0812	0.802	0.0636
80%	0.468	9.27×10^{-5}	0.742	0.0621	0.76	0.0449
90%	0.437	9.27×10^{-5}	0.640	0.0403	0.699	0.0300

Table 5.7: ROC contingency table for the $t + 5$ GPNN model.

5-hr-ahead prediction						
	Super Storm		*Intense Storm*		*Moderate*	
Threshold	**TPR**	**FPR**	**TPR**	**FPR**	**TPR**	**FPR**
10%	0.812	3.06×10^{-3}	0.956	0.316	0.962	0.346
20%	**0.812**	$\mathbf{1.02 \times 10^{-3}}$	0.934	0.246	0.945	0.265
30%	0.750	4.63×10^{-4}	0.917	0.189	0.926	0.215
40%	0.719	9.27×10^{-5}	0.891	0.148	0.906	0.171
50%	0.625	9.27×10^{-5}	**0.856**	**0.120**	0.881	0.139
60%	0.562	9.27×10^{-5}	0.824	0.0942	**0.853**	**0.107**
70%	0.468	0	0.779	0.0740	0.810	0.081
80%	0.468	0	0.725	0.055	0.754	0.0654
90%	0.468	0	0.639	0.0381	0.685	0.0430

Table 5.8: ROC contingency table for the $t + 6$ GPNN model.

6-hr-ahead prediction						
	Super Storm		*Intense Storm*		*Moderate*	
Threshold	**TPR**	**FPR**	**TPR**	**FPR**	**TPR**	**FPR**
10%	**0.500**	$\mathbf{8.34.10^{-3}}$	0.953	0.352	0.932	0.307
20%	0.437	4.92×10^{-3}	0.928	0.289	0.909	0.241
30%	0.437	3.24×10^{-3}	0.904	0.244	0.886	0.186
40%	0.406	2.78×10^{-3}	0.890	0.202	0.862	0.161
50%	0.375	1.76×10^{-3}	**0.859**	**0.167**	**0.834**	**0.130**
60%	0.375	1.39×10^{-3}	0.821	0.138	0.798	0.113
70%	0.281	7.47×10^{-4}	0.788	0.115	0.757	0.0914
80%	0.281	3.70×10^{-4}	0.735	0.0926	0.712	0.0693
90%	0.281	2.78×10^{-4}	0.661	0.0691	0.649	0.0455

Table 5.9: Storm events used to evaluate GPNN model

Start Date	Start Time	End Date	End Time	min. Dst
2001/03/19	15:00	2001/03/21	23:00	$-149\,\mathrm{nT}$
2001/03/31	04:00	2001/04/01	21:00	$-387\,\mathrm{nT}$
2001/04/18	01:00	2001/04/18	13:00	$-114\,\mathrm{nT}$
2001/04/22	02:00	2001/04/23	15:00	$-102\,\mathrm{nT}$
2001/08/17	16:00	2001/08/18	16:00	$-105\,\mathrm{nT}$
2001/09/30	23:00	2001/10/02	00:00	$-148\,\mathrm{nT}$
2001/10/21	17:00	2001/10/24	11:00	$-187\,\mathrm{nT}$
2001/10/28	03:00	2001/10/29	22:00	$-157\,\mathrm{nT}$
2002/03/23	14:00	2002/03/25	05:00	$-100\,\mathrm{nT}$
2002/04/17	11:00	2002/04/19	02:00	$-127\,\mathrm{nT}$
2002/04/19	09:00	2002/04/21	06:00	$-149\,\mathrm{nT}$
2002/05/11	10:00	2002/05/12	16:00	$-110\,\mathrm{nT}$
2002/05/23	12:00	2002/05/24	23:00	$-109\,\mathrm{nT}$
2002/08/01	23:00	2002/08/02	09:00	$-102\,\mathrm{nT}$
2002/09/04	01:00	2002/09/05	00:00	$-109\,\mathrm{nT}$
2002/09/07	14:00	2002/09/08	20:00	$-181\,\mathrm{nT}$
2002/10/01	06:00	2002/10/03	08:00	$-176\,\mathrm{nT}$
2002/11/20	16:00	2002/11/22	06:00	$-128\,\mathrm{nT}$
2003/05/29	20:00	2003/05/30	10:00	$-144\,\mathrm{nT}$
2003/06/17	19:00	2003/06/19	03:00	$-141\,\mathrm{nT}$
2003/07/11	15:00	2003/07/12	16:00	$-105\,\mathrm{nT}$
2003/08/17	18:00	2003/08/19	11:00	$-148\,\mathrm{nT}$
2003/11/20	12:00	2003/11/22	00:00	$-422\,\mathrm{nT}$
2004/01/22	03:00	2004/01/24	00:00	$-149\,\mathrm{nT}$
2004/02/11	10:00	2004/02/12	00:00	$-105\,\mathrm{nT}$
2004/04/03	14:00	2004/04/04	08:00	$-112\,\mathrm{nT}$
2004/07/22	20:00	2004/07/23	20:00	$-101\,\mathrm{nT}$
2004/07/24	21:00	2004/07/26	17:00	$-148\,\mathrm{nT}$
2004/07/26	22:00	2004/07/30	05:00	$-197\,\mathrm{nT}$
2004/08/30	05:00	2004/08/31	21:00	$-126\,\mathrm{nT}$
2004/11/11	22:00	2004/11/13	13:00	$-109\,\mathrm{nT}$
2005/01/21	18:00	2005/01/23	05:00	$-105\,\mathrm{nT}$
2005/05/07	20:00	2005/05/09	10:00	$-127\,\mathrm{nT}$
2005/05/29	22:00	2005/05/31	08:00	$-138\,\mathrm{nT}$
2005/06/12	17:00	2005/06/13	19:00	$-106\,\mathrm{nT}$
2005/08/31	12:00	2005/09/01	12:00	$-131\,\mathrm{nT}$
2006/04/13	20:00	2006/04/14	23:00	$-111\,\mathrm{nT}$
2006/12/14	21:00	2006/12/16	03:00	$-147\,\mathrm{nT}$
2011/09/26	14:00	2011/09/27	12:00	$-101\,\mathrm{nT}$
2011/10/24	20:00	2011/10/25	14:00	$-132\,\mathrm{nT}$
2012/03/08	12:00	2012/03/10	16:00	$-131\,\mathrm{nT}$
2012/04/23	11:00	2012/04/24	13:00	$-108\,\mathrm{nT}$
2012/07/15	01:00	2012/07/16	23:00	$-127\,\mathrm{nT}$
2012/09/30	13:00	2012/10/01	18:00	$-119\,\mathrm{nT}$
2012/10/08	02:00	2012/10/09	17:00	$-105\,\mathrm{nT}$
2012/11/13	18:00	2012/11/14	18:00	$-108\,\mathrm{nT}$
2013/03/17	07:00	2013/03/18	10:00	$-132\,\mathrm{nT}$
2013/05/31	18:00	2013/06/01	20:00	$-119\,\mathrm{nT}$
2014/02/18	15:00	2014/02/19	16:00	$-112\,\mathrm{nT}$

Chapter 6

Identifying Radiation Belt Parameters: A Bayesian Approach

We present a novel method, which assimilates sparse irregular observations of a field with its governing physical dynamics. Our method uses a basis function approach coupled with a least squares support vector machine objective function which gives different weights to errors arising due to data fitting and satisfaction of physical constraints. The method is applicable to linear PDE systems, it incorporates physical models into classical least squares techniques for the purpose of data assimilation and uncertainty quantification of latent parameters. We apply this method to the problem of identifying radiation belt parameters from sparse observations.

* * *

This chapter is based on research which is in preperation for publication. Research led by M. Chandorkar in collaboration with R. Sarma and supervised by E. Camporeale. Y. Sphrits and A. Drozdov provided assistance in processing Van Allen probe data.

6.1 Introduction

The Earth's radiation belts are the regions of space near the Earth that extend between $2R_E$ and $8R_E$ ($R_E = 6372\,\mathrm{km}$, the radius of the Earth), where the terrestrial magnetic field traps electrons and ions in complex electromagnetic orbits [Van Allen and Frank, 1959]. Since their discovery, the belts have been the subject of intensive research due to their complex behavior and damaging effects on spacecraft [Gubby and Evans, 2002, Welling, 2010, Baker, 2002].

Radiation belt particles generally execute three types of periodic motion, each with its own corresponding adiabatic invariant: gyration around magnetic field lines, bounce along field lines, and drift around the Earth. During active times, when conditions change on time scales shorter than the periods of motion, adiabaticity can be broken and particle motion can not be simply decomposed into the aforementioned components. In this case, particle motion can be represented diffusively along each component via the *Fokker-Planck* equation yielding a useful model of radiation belt dynamics [Schulz and Lanzerotti, 1974].

The third invariant represents the total magnetic flux enclosed within a full particle orbit. Generally, its normalised inverse, the well known Roederer L^* [Roederer, 1970, ch 3] which is analogous to radial distance from the center of the Earth (in Earth radii) to the equatorial crossing point of a bouncing particle, is used when modeling radiation belt dynamics. Diffusion in L^* alone (the other invariants shall be considered conserved) accounts for the capture and inward radial transport of radiation belt particles [Fälthammar, 1965, Roederer, 1970].

One of the main difficulties of using a physics-based model for studying and forecasting energetic electrons in the radiation belt is that the parameters that characterise the Fokker-Planck equation, namely diffusion tensor and loss term, are not directly observable. Hence, their determination is an *inverse problem*, which is generally difficult to solve and can often become ill-posed.

In this chapter we propose an inference model which can learn from sparse data while taking into account prior knowledge of the system dynamics in the form of a linear partial differential equation. The method replaces a standard finite difference solver with a surrogate model which tries to fit the observations and the system dynamics. The surrogate is expressed as a basis function expansion whose coefficients are computed by formulating a *least squares support vector machine* (LSSVM) like optimisation objective.

In the proceeding sections, we give a short introduction to the radial diffusion equation used in magnetospheric physics. After an overview of the parameterizations of the radial diffusion unknowns used by the research community, we give a detailed formulation of our proposed method and demonstrate how it can be used for performing inference over said diffusion parameters.

6.2 Radial Diffusion

As discussed in section 6.1 above, plasma motion in the radiation belt can be modeled as a simplified one dimensional version of the Fokker-Planck equation. The resulting system, shown in equation (6.1), known as *radial diffusion* in radiation belt physics [Lyons and Schulz, 1989]. It governs the time evolution of the *phase space density* (PSD) of particles, which is expressed as a function of the Roederer L^* (henceforth denoted as ℓ), and the time coordinate t.

$$\frac{\partial f}{\partial t} = \ell^2 \frac{\partial}{\partial \ell}\left(\frac{\kappa(\ell,t)}{\ell^2}\frac{\partial f}{\partial \ell}\right)_{\mathcal{M},J} - \lambda(\ell,t)f + q(\ell,t) \tag{6.1}$$

The key quantities in the radial diffusion system are summarised below.

1. f: The PSD, for fixed values of the first and second adiabatic invariants $\mathcal{M}$ and $\mathcal{J}$ respectively, as a function of space (ℓ) and time (t).

2. $\kappa(\ell,t)$: A space and time varying diffusion field.

3. $\lambda(\ell,t)$: The particle loss rate, a non-negative quantity which indicates how quickly particles are lost from the radiation belts, due to mechanisms other than radial diffusion.

4. $q(\ell,t)$: An optional particle injection rate or source term. If this term is omitted (i.e. $q(\ell,t) = 0$), then the boundary conditions $f(\ell_{min},t)$

and $f(\ell_{max}, t)$ must be also specified along with the initial condition $f(\ell, 0)$ to solve equation (6.1).

Interested readers can refer to section 3.3.1 for a detailed explanation on plasma diffusion, adiabatic invariants and plasma motion in the Earth's radiation belts.

6.2.1 Diffusion Parameters

To solve the radial diffusion system (equation (6.1)), the quantities $\kappa(\ell, t)$, $\lambda(\ell, t)$ and $q(\ell, t)$ need to be specified. It is a common practice [see Selesnick et al., 1997, Brautigam and Albert, 2000, Fei et al., 2006 and Shprits et al., 2007] to parametrize the diffusion field κ and loss rate λ as

$$\kappa(\ell, t), \lambda(\ell, t) \sim \alpha \ell^{\beta} 10^{b\mathrm{Kp}(t)}. \tag{6.2}$$

The quantities α, β, and b above are parameters which define the diffusion field and loss rate. The quantity $\mathrm{Kp}(t)$ is the well known Kp index, a proxy for the global geomagnetic activity [Bartels et al., 1939]. In equation (6.3) we propose a parametrization of the source term q which can approximate particle injection through the upper boundary (ℓ_{max}) of the radiation belt.

$$q(\ell, t) \sim \exp(\alpha - \beta(\ell - \ell_{max})^2) 10^{b\mathrm{Kp}(t)} \tag{6.3}$$

The expression above is similar to the formulations of the diffusion and loss terms presented earlier, however it is distinguished by its rapid spatial decay away from the upper boundary ℓ_{max}.

It is possible to impose reasonable restrictions on the domains of parameters α, β, and b of the source term $q(\ell, t)$. The parameter α can take values on the entire real number line, but since we expect $q(\ell, t)$ to model particle injection from the outer boundary ℓ_{max} during active geomagnetic conditions, β and b must both positive. We shall see in section 6.5 that these constraints dictate the class of prior distributions can be chosen for α, β, and b, and alleviate problems of identifiability.

6.3 PDE Inverse Problems

Partial differential equations are usually solved by approximating the derivatives involved on a spatiotemporal grid with numerical quadrature. Solving of a PDE system is referred to as the *forward problem*, therefore inference of the PDE parameters from observations is naturally called the *inverse problem*.

The PDE constrained inverse problem is defined as follows: given a set of noisy observations $\mathcal{D} = \{(x_i, y_i) : x \in \mathcal{X} \times [0, \infty), y \in \mathbb{R}\}$ of a physical quantity $f(x)$ which is governed by the differential equation

$$\mathcal{L}_\theta f(x) = q_\theta(x), \tag{6.4}$$

estimate the parameters θ of the forward model $\mathcal{F} = (\mathcal{L}_\theta, q_\theta)$, where $\mathcal{X} \times [0, \infty)$ is the spatiotemporal domain, $\mathcal{L}_\theta$ is a differential operator, and q_θ is a *source term*.

For the radial diffusion system shown in equation (6.1), the differential operator $\mathcal{L}_\theta$ is defined as

$$\mathcal{L}_\theta = \frac{\partial}{\partial t} - \ell^2 \frac{\partial}{\partial \ell} \left(\frac{\kappa(\ell, t)}{\ell^2} \frac{\partial}{\partial l} \right) + \lambda(\ell, t).$$

In this case θ is a collection of parameters which would specify closed form expressions for $\kappa(\ell, t)$, $\lambda(\ell, t)$ and $q(\ell, t)$ such as equations (6.2) and (6.3).

6.3.1 Bayesian PDE Inverse Problem

Bayesian statistics [Lee, 1997] treats the problem of parameter determination as a problem of probabilistic inference. By specifying: 1. a prior probability distribution over the system parameters, and 2. a likelihood distribution which determines the conditional distribution of the observations given the parameters, the *posterior distribution* or the conditional probability distribution of the parameters given the observations can be computed using Bayes rule.

In PDE constrained inverse problems, the object of interest is the distribution $p(\theta|\mathcal{D}, \mathcal{F})$ which is given by

$$p(\theta|\mathcal{D}, \mathcal{F}) = \frac{p(\mathcal{D}|\theta, \mathcal{F}) p(\theta)}{p(\mathcal{D}|\mathcal{F})},$$

where $p(\theta)$ is the prior distribution, $p(\mathcal{D}|\theta, \mathcal{F})$ is the likelihood of the data in the forward model (the radial diffusion PDE), and the denominator

$$p(\mathcal{D}|\mathcal{F}) = \int p(\mathcal{D}|\theta, \mathcal{F}) p(\theta) d\theta$$

is a normalisation term known as the *model evidence*. For all but a few simple problems, the model evidence term cannot be computed in closed form, therefore, the posterior probability can only be computed up to a normalisation factor. Markov Chain Monte Carlo (MCMC) based methods are used to generate samples from the posterior distribution, as they work with probability density ratios and hence do not require expensive numerical integrations needed to compute $p(\mathcal{D}|\mathcal{F})$.

In PDE constrained Bayesian inverse problems, computation of the likelihood $p(\mathcal{D}|\theta, \mathcal{F})$ requires solving the forward problem specified in equation (6.4). Real world inverse problems often involve observations which are sparse and irregularly spaced, therefore finite difference based PDE solution methods must be combined with interpolation in order to compute the observation likelihood. Mesh-free PDE solution methods provide are an alternative to finite difference methods where the PDE solution can be computed on sets of arbitrarily spaced domain points.

6.3.2 Related Work

A large body of research has been devoted to the development of mesh-free methods for PDEs; they have been applied in both the forward and inverse problems. Below we summarise some prominent themes in mesh-free methods.

Gaussian process (GP) models [Rasmussen and Williams, 2005] have a rich theory which has much overlap with linear systems and deterministic and stochastic differential equations. Skilling [1992] presented one of the earliest works which focused on calculating solutions of ordinary differential equations (ODE) systems with Gaussian Process methodology, Graepel [2003] applied it for solving linear partial differential equations with Dirichlet and Von Neumann boundary conditions.

Interplay between linear operators and GP models applied to Bayesian filtering was investigated by Särkkä [2011]. Dondelinger et al. [2013] proposed an adaptive gradient matching technique to used Gaussian Process models for inferring parameters of coupled ODE systems.

Raissi et al. [2018] proposed the *numerical Gaussian process* methodology which placed GP priors on spatial fields and quantified uncertainties in the solutions of time discretised PDE systems.

Neural networks were also employed for solutions of boundary value problems in the works such as Lagaris et al. [1998], Aarts and van der Veer [2001], Tsoulos et al. [2009], Baymani et al. [2011] which used feedforward networks for calculating solutions to the Stokes problem. These approaches generally revolved around decomposing the solution into two components, i.e. the first one satisfying the boundary conditions and the second one represented by the feedforward network.

Raissi [2018] used neural networks for the identification of PDE systems from observations of a spatiotemporal field. The proposed method was composed of two neural networks, the first one which proposed a candidate PDE built from a predefined dictionary of derivative terms, and the second network which computed an approximate solution of the PDE constructed by the first.

Radial basis functions (RBF) were first applied for solution of PDE problems in Kansa [1990], the authors used colocation with *multi-quadric* basis functions for approximating solutions of boundary value problems.

Radial basis functions have been applied for the mesh-free solutions of Poisson PDE systems [Aminataei and Mazarei, 2008, Duan, 2008, Duan and Tan, 2006, Elansari et al., 2001], as well as the Poisson control problem [Pearson, 2013]. Further applications of RBFs include atmospheric flow [Tillenius et al., 2015], convection-diffusion [Safdari-Vaighani et al., 2015] and Schrödinger's equation [Kormann and Larsson, 2013]. Refer to Fornberg and Flyer [2015] for a recent textbook with geoscience applications.

Least squares support vector machines have also been applied to calculating approximate solutions to PDEs as well as parameter estimation of delay differential equations (DDE) [Mehrkanoon and Suykens, 2015, 2012, Mehrkanoon et al., 2014]. The approach adopted in Mehrkanoon et al. [2014] expressed the parameter estimation of the time delay in a DDE as an algebraic optimisation problem resulting in closed-form approximation for the time varying parameters while avoiding iterative simulation of the dynamical system (governed by the delay differential equations) in the parameter estimation process.

Probabilistic numeric methods (PNM), an area which concerns with the quantification of errors and uncertainties in numerical methods arising from loss of precision due to limitations of time and hardware [Hennig et al., 2015]. Applications of PNM range from Bayesian quadrature, optimisation, mesh-free solutions of PDEs, and PDE constrained Bayesian inverse problems.

Conrad et al. [2017] propose a probabilistic time integrator for quantifying probability measures over solutions of ODE systems. Cockayne et al. [2017] introduce a probabilistic mesh-free method (PMM) for quantifying uncertainty over the solution space of Linear PDEs, their model consists of a GP prior which is conditioned on a finite set of design points (or colocation points) which constrain the GP based on the PDE dynamics and its boundary conditions. The authors apply the PMM method for quantifying uncertainties in the forward as well as the inverse problem, and also provide theoretical results regarding the rates of convergence of the posterior distribution in both cases.

6.4 Methodology

Our approach solving the radial diffusion inverse problem proceeds has two components.

1. A surrogate PSD model built by formulating a modified version of the least squares support vector machine predictor for obtaining a closed form approximation to f which tries to satisfy equation (6.1) on a fixed set of *colocation* points while minimizing error on a set of sparse noisy observations.

2. A multivariate Gaussian likelihood over the noisy observations, the mean and covariance of which are computed using the surrogate model.

Below we describe the formulation of the surrogate model for the PSD in radial diffusion.

6.4.1 Phase Space Density Surrogate

Let $\mathcal{D} = \{(x_i^o, y_i) : i = 1 \cdots n_o\}$ be a set of noisy observations of the phase space density f, where $x_i = (\ell_i, t_i)$ are points in the space-time domain $[\ell_{min}, \ell_{max}] \times [0, \infty)$. We seek a linear estimator for f of the form $\hat{f}(x) = w^T \varphi(x) + b$, where $\varphi(.) : \mathbb{R}^2 \rightarrow \mathbb{R}^d$ is a d dimensional feature map and b is a scalar intercept.

Further let $\mathcal{C} = \{(x_i^c, q_i) : i = 1 \cdots n_c\}$ be a set of colocation points on which we aim to enforce radial diffusion dynamics. The values q_i represent the particle injection rate at x^c, and are calculated using equation (6.3).

We exploit the linearity of the differential operator $\mathcal{L}_\theta$ and note that $\mathcal{L}_\theta[\hat{f}(x)] = w^T \mathcal{L}_\theta[\varphi(x)] + \mathcal{L}_\theta[b]$, yielding an estimator $\hat{q}(x) = w^T \psi(x) + \mathcal{L}_\theta[b]$ where $\psi_\theta(x) = \mathcal{L}_\theta[\varphi(x)]$. Determining $w \in \mathbb{R}^d$ can now be cast as the following constrained L_2 regularised least squares problem.

$$\begin{aligned}
&\min_{w,e,\epsilon} \ \mathcal{J}(w, e, \epsilon; \theta) = \\
&\frac{1}{2} w^T w + \frac{1}{2\gamma_o} \sum_{k=1}^{n_o} e_k^2 + \frac{1}{2\gamma_c} \sum_{k=1}^{n_c} u_k \epsilon_k^2 \\
s.t& \\
&y_i = w^T \varphi(x_i^o) + b + e_i, \quad i = 1 \cdots n_o \\
&q_i = w^T \psi_\theta(x_i^c) + \mathcal{L}_\theta(b) + \epsilon_i, \quad i = 1 \cdots n_c
\end{aligned} \tag{6.5}$$

The quantities γ_o and γ_c are weights attached to the errors on observations and colocation points respectively. Thus by smoothly varying them we can assign higher or lower importance to the surrogate model, in order to fit the observational data and the dynamics of the physical system. The quantities u_i enable us to weigh each colocation point differently.

It can be seen that system in equation (6.5) is similar to the formulation of the LSSVM model, while incorporating the dynamics of linear PDE systems into its loss function. Solving the system in equation (6.5) is achieved by constructing its *Lagrangian* given in equation (6.6).

$$\begin{aligned}
&\mathfrak{L}(w, e, \epsilon, \alpha_{1\cdots k}, \beta_{1\cdots k}; \theta; \gamma_o; \gamma_c) = \\
&\frac{1}{2} w^T w + \frac{1}{2\gamma_o} \sum_{k=1}^{n_o} e_k^2 + \frac{1}{2\gamma_c} \sum_{k=1}^{n_c} u_k \epsilon_k^2 \\
&+ \sum_{k=1}^{n_o} \alpha_k (y_k - w^T \varphi(x_k^o) - b - e_k) \\
&+ \sum_{k=1}^{n_c} \beta_k (q_j - w^T \psi_\theta(x_j^c) - \mathcal{L}_\theta[b] - \epsilon_j)
\end{aligned} \tag{6.6}$$

The quantities $\alpha_1, \cdots, \alpha_{n_o}$ and $\beta_1, \cdots, \beta_{n_c}$ are the *Lagrange multipliers* introduced for equality constraints of the system. Applying the *Karush-Kuhn-Tucker* (KKT) conditions [Karush, 1939, Kuhn and Tucker, 1951], the solution of the optimisation problem in equation (6.5) can be expressed in terms of the Lagrange multipliers $\alpha = (\alpha_1, \cdots, \alpha_{n_o})$ and $\beta = (\beta_1, \cdots, \beta_{n_c})$.

$$\begin{bmatrix} 0 & \mathbf{1}^T & \mathbf{\Lambda}_\theta{}^T \\ \mathbf{1} & \Omega + \gamma_o I & \Omega_* \\ \mathbf{\Lambda}_\theta & \Omega_*^T & \Omega_{**} + \gamma_c U \end{bmatrix} \begin{bmatrix} b \\ \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} 0 \\ y \\ q \end{bmatrix} \tag{6.7}$$

The components of the symmetric block matrix system on the left hand side of equation (6.7) are

- $\Omega \in \mathbb{R}^{n_o \times n_o} : \omega_{ij} = \varphi(x_i^o)^T \varphi(x_j^o)$,
- $\Omega_{**} \in \mathbb{R}^{n_c \times n_c} : \omega_{ij}^{**} = \psi(x_i^c)^T \psi(x_j^c)$,
- $\Omega_* \in \mathbb{R}^{n_o \times n_c} : \omega_{ij}^{*} = \varphi(x_i^o)^T \psi(x_j^c)$ and
- $U \in \mathbb{R}^{n_c \times n_c} = \begin{bmatrix} u_1 & \cdots & 0 \\ \vdots & \ddots & 0 \\ 0 & \cdots & u_{n_c} \end{bmatrix}.$
- $\Lambda_\theta \in \mathbb{R}^{n_c} : \lambda_i = \lambda_\theta(x_i^c)$

The surrogate model (equation (6.8)) can now be used to estimate the phase space density at a point $x = (\ell, t)$.

$$\hat{f}(x;\theta) = \sum_{k=1}^{n_o} \alpha_k \varphi(x)^T \varphi(x_k^o) + \sum_{k=1}^{n_c} \beta_k \varphi(x)^T \psi_\theta(x_k^c) + b \tag{6.8}$$

Choice of $\varphi(.)$

The function $\varphi(.)$ maps the spatiotemporal input x to a d dimensional feature space, which is subsequently the input for the family of linear surrogate predictors $\hat{f}$. It is therefore natural to express $\varphi(.)$ as a basis function expansion. The choice of $\varphi(.)$ influences the continuity characteristics and representational capability of the surrogate model class $\hat{f}$.

There exist several options regarding the choice of basis, orthogonal polynomials, Fourier series, radial basis functions, etc. We express $\varphi(.)$ as a product of space and time bases $\phi(.)$ and $\varpi(.)$ respectively, as follows:

$$\varphi_{i,j}(\ell, t) = T_i \left(2\frac{\ell - \ell_{min}}{\ell_{max} - \ell_{min}} - 1 \right) \varpi_j(t) \ ,$$

where $T_i(.)$ is the Chebyshev polynomial of the first kind, of degree i, and $\varpi_j(t) = \frac{1}{\sqrt{1+(t-t_i)^2/s_i^2}}$ is the inverse multi-quadric RBF centered at t_i and having length scale s_i.

Role of γ_o, γ_c and u_i

The quantities γ_o and γ_c serve to control the importance assigned to errors made on the observations and colocation points respectively. Varying them gives the modeler the ability to vary the behavior of the surrogate model. In the limiting case of γ_c tending to zero, the model behaves as if the PDE dynamics is enforced as a hard constraint. This case is equivalent to the following formulation.

$$\begin{aligned} \min_{w,e} \ \mathcal{J}(w, e, \epsilon; \theta) &= \frac{1}{2} w^T w + \frac{1}{2\gamma_o} \sum_{k=1}^{n_o} e_k^2 \\ & s.t \\ & y_i = w^T \varphi(x_i^o) + b + e_i, \quad i = 1 \cdots n_o \\ & q_i = w^T \psi_\theta(x_i^c) + \mathcal{L}_\theta(b), \quad i = 1 \cdots n_c \end{aligned} \tag{6.9}$$

Although choosing $\gamma_c = 0$ is an appropriate choice if the physical dynamics need to be enforced as a constraint, it can possibly lead to numerical instabilities in inverting system in equation (6.7) and hence choosing a non zero value for γ_c works better in practice.

The weights u_i have a special interpretation in the context of equation (6.6). It is possible to interpret the term $\sum_{k=1}^{n_c} u_k \epsilon_k^2$ as a quadrature approximation to the integrated error of the surrogate model with respect to the governing dynamics $\int_{x \in \mathcal{D}} ||\mathcal{L}_\theta[\hat{f}(x)] - Q(x)||^2$. When $u_i = \frac{1}{n_c}$, this corresponds to the *Monte Carlo* quadrature of the integrated error, but it is possible to improve the quadrature accuracy by using *Gauss-Legendre* quadrature.

In section 6.5, we use eight point Gauss-Legendre quadrature in space and time dimension each, thereby setting $n_c = 64$ and weights u_i to appropriate values as dictated by the chosen quadrature rule [Abramowitz and Stegun, 1972].

6.4.2 Quantifying Observation Likelihood

We assume a multivariate Gaussian distribution (equations (6.10) to (6.13)) for calculating the likelihood of the observations conditioned on the system parameters θ.

$$\mathbf{y}|x_1, \cdots, x_{n_o}, \theta \sim \mathcal{N}(\mu_f, \Sigma) \tag{6.10}$$

$$\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_{n_o} \end{bmatrix} \tag{6.11}$$

$$\mu_f = \begin{bmatrix} \hat{f}(x_1) \\ \vdots \\ \hat{f}(x_{n_o}) \end{bmatrix} \tag{6.12}$$

$$\Sigma = \begin{bmatrix} K(x_1, x_1) & \cdots & K(x_1, x_{n_o}) \\ \vdots & \ddots & \vdots \\ K(x_{n_o}, x_{n_1}) & \cdots & K(x_{n_o}, x_{n_o}) \end{bmatrix} \tag{6.13}$$

The surrogate model (equation (6.8)) gives the mean value for the phase space density, and we use a hybrid RBF covariance function

$$K(x_i, x_j) = \sigma^2 \exp\left(-\frac{1}{2}\frac{|\ell_i - \ell_j|^2}{s} + \frac{|t_i - t_j|}{r}\right) + \varphi(x_i)^T \varphi(x_j)$$

to quantify the covariance of the phase space density f over two points $x_i = (\ell_i, t_i)$ and $x_j = (\ell_j, t_j)$ in the domain.

6.4.3 Inference

We employ the adaptive Metropolis algorithm as proposed by Haario et al. [2001], for sampling system parameters. The adaptive Metropolis algorithm adapts the exploration variance according to the running sample statistics of the MCMC procedure.

Table 6.1: Parameters: Prior Distributions

Parameter	Prior
α	Uniform$(-10, 10)$
β	Uniform$(0, 10)$
b	Uniform$(0, 2)$

6.5 Experiments

We tested our proposed model on two experiments discussed below.

1. **Synthetic Problem**: Using a synthetically generated data set, we evaluated the ability of our model to identify radial diffusion parameters, when their ground truth values are known.

2. **Radiation Belt Particle Injection**: Our model was used to quantify uncertainties in the source term parameters from Van Allen probe data. The source term $q(\ell, t)$ (equations (6.1) and (6.3)) models particle injection in the radiation belts through their outer boundary.

In both experiments, we chose $\gamma_o = 2$ and $\gamma_c = 10^{-4}$. The parameters of $\kappa(\ell, t)$ and $\lambda(\ell, t)$ were fixed to the values computed in Brautigam and Albert [2000] and Shprits et al. [2007] respectively (see table 6.2) while inference was performed only on the parameters of $q(\ell, t)$.

In both experiments, the bases were chosen as described in section 6.4.1. In the synthetic problem, we chose 5 basis functions in space and 20 basis functions in time respectively; however in the van allen probe problem, we chose 6 basis functions in space and 500 in time. With regards to the temporal basis, the RBF nodes were placed on equally spaced grid and their length scales were set to the grid cell size.

The quantities s and r which are the length scales of the covariance function can also be treated as system parameters which can be sampled by the inference procedure. Since the core aims of this research was the quantification of the uncertainty over the parameters of the radial diffusion system, we treated the covariance function parameters as fixed.

The prior distributions chosen for the parameters are shown in table 6.1, they are in accordance with the constraints discussed in section 6.2.1. The posterior distribution over the parameters of $q(\ell, t)$ was sampled via the adaptive Metropolis algorithm, the first 2000 samples were discarded as the 'burn in' of the Markov Chain. A total of 2000 samples were generated after the burn in period, and used for the reporting of results.

After sampling from the posterior distributions, the samples obtained are taken together with 2000 samples drawn from the prior distributions, and visualised in the following charts.

1. *Density estimate charts*: These plots show smoothed kernel density estimates computed from samples drawn from the prior and posterior distributions.

2. *Scatter charts*: The prior and posterior samples are plotted together along the parameter pairs α, b and β, b.

Table 6.2: Parameters: Ground Truth

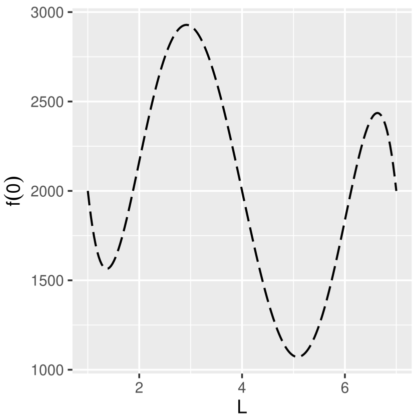
Quantity	α	β	b	Reference
q	-1	2.5	0.75	N.A.
κ	$\log(4.731 \times 10^{-10})$	10	0.506	Brautigam and Albert [2000]
λ	$\log(0.3678)$	0.5	-0.2	Shprits et al. [2007]

6.5.1 Synthetic Data

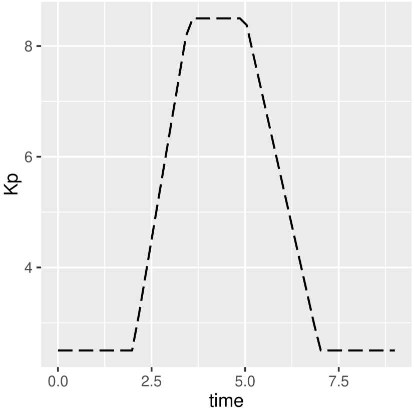
We generated a noisy synthetic data set by solving the radial diffusion equation using a forward difference algorithm. The task of our model was to use the observations to infer the parameters α, β, and b of the source term $q(\ell, t)$ that were used to generate the observations.

The ground truth values of the radial diffusion parameters are listed in table 6.2. The initial phase space density $f(t = 0)$ was chosen as follows (see figure 6.1a):

$$f(t = 0) = 2000 + 500(T_3(\ell_*) - T_5(\ell_*))$$
$$\ell_* = 2\frac{\ell - \ell_{min}}{\ell_{max} - \ell_{min}} - 1 \ ,$$



(a) The initial condition $f(\ell, 0)$.



(b) The time evolution of the Kp index.

Figure 6.1: Synthetic data generation.

where $T_n(.)$ is the is the Chebyshev polynomial of degree n. The evolution of the Kp index was assumed to be an idealised version of a geomagnetic storm (see figure 6.1b) and defined as

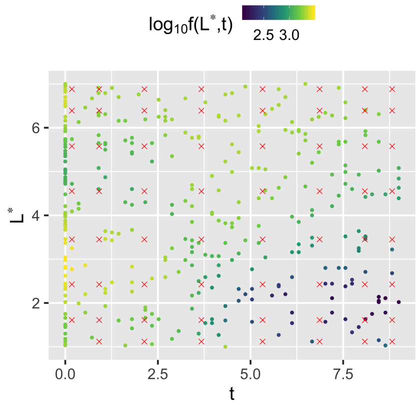
$$\mathrm{Kp}(t) = \begin{cases} 2.5 & 0 \leq t < 2 \\ 2.5 + 4(t - 2d) & 2 \leq t < 3.5 \\ 8.5 & 3.5 \leq t < 5 \\ 23.5 - 3t & 5 \leq t < 7 \\ 2.5 & t \geq 7 \end{cases} .$$

The radial diffusion solver was run for domain limits $\ell \in [1, 7], t \in [0, 9]$ with 200 bins in the spatial and 50 bins in the temporal domains respectively. After the solution f was computed, points was sub-sampled uniformly such that 250 points lying in the interior of the domain, and 50 points at the initial time step ($t = 0$) were selected. These observations were then perturbed by Gaussian noise to yield the final observation set $\mathcal{D}$ which was provided to the surrogate model $\hat{f}(x)$. The generated data set is plotted in figure 6.2a.

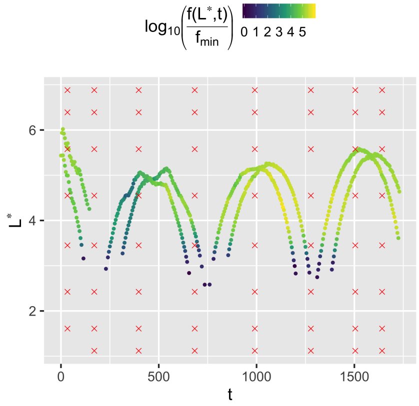
6.5.2 Radiation Belt Data: Van Allen Probes

Van Allen probe data from the *MagEIS* instrument for the time period ranging from $6:00$ 17^{th} March 2013 to $11:00$ 18^{th} March 2013 was extracted and used for performing inference over the parameters of $q(\ell, t)$. This particular period was chosen because active geomagnetic conditions occurred in it. Figure 6.2b plots the Van Allen PSD data on a logarithmic scale. To avoid dealing with extremely small values, we use $\log_{10}\left(\frac{f(\ell,t)}{f_{min}}\right)$, the logarithm of the ratio of the PSD to its observed minimum value in the data.

There are two differences which stand out between the van allen data set and the synthetic data used in section 6.5.1 above. Firstly, the synthetic data is randomly sampled while the Van Allen probe data is sampled along their orbits (there are two probes). Secondly, the synthetic data is has a larger density of observations in the space time domain as compared to the Van Allen data. These factors among others influence the uncertainties in the posterior inference, which is seen in section 6.5.3 below.

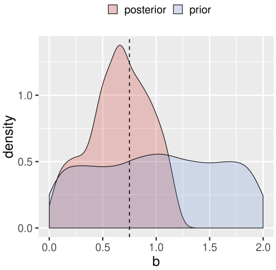


(a) Synthetic data used in section 6.5.1. Red crosses indicate colocation points chosen according to 8 point Gauss-Legendre quadrature in space and time.

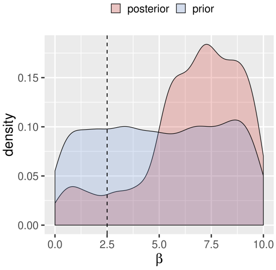


(b) Van Allen probe data used in section 6.5.2. Time is measured in minutes from the starting point $6:00$ 17th March 2013. Red crosses indicate colocation points chosen according to 8 point Gauss-Legendre quadrature in space and time.

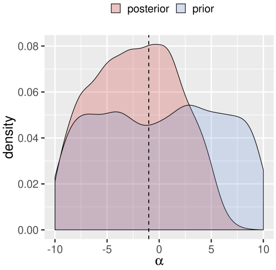
Figure 6.2: Data sets used in the experiments.



(a) Parameter b



(b) Parameter β



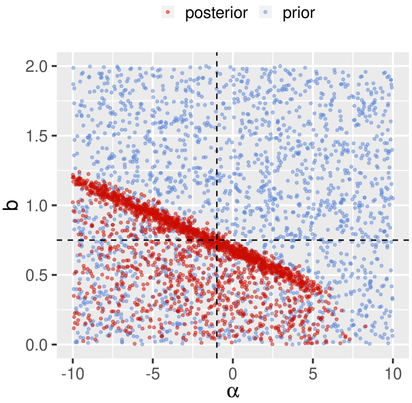
(c) Parameter α

Figure 6.3: **Synthetic Data**: Comparing prior and posterior densities for parameters of $q(\ell, t)$, the black dotted line indicates the ground truth.

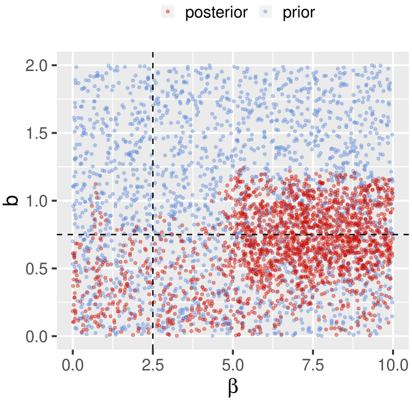
6.5.3 Results

Synthetic Problem

For the synthetic problem, the inferred posterior distribution for parameters of the particle injection rate $q(\ell, t)$ are shown in figures 6.3a to 6.3c. We see



(a) Scatter chart: α versus b.



(b) Scatter chart: β versus b.

Figure 6.4: **Synthetic Data**: Prior and posterior samples drawn from parameters of $q(\ell, t)$.

that the marginal posterior distributions of b and α have high probability density near the ground truth.

The scatter charts in figures 6.4a and 6.4b help us to identify regions in the parameter space to which our model assigns high probability. From the scatter plot in figure 6.4a, we see a clear negative correlation between α and b. This negative correlation is a natural consequence of the parametrization of $q(\ell, t)$ in equation (6.3). When α is increased, b must appropriately decreased if one is to explain the PSD observations. Examining the dependence between α and b in figure 6.4a, one can make the interpretation that along the red linear region corresponding to high posterior probability, all values of α and b can explain the PSD observations with similar likelihood. We also observe that the density of posterior samples is greater below the red linear region as compared to above it.

From figures 6.3b and 6.4b, we observe that the model doesn't identify the parameter β. Its posterior probability distribution does not have significantly reduced uncertainty compared to its prior. The differences in the inference results between parameters α, b, and β can be attributed to the vast differences in the sensitivity of the radial diffusion solution to α, b, and β. The interested reader can refer to appendix B for an introduction to the sensitivity analysis of partial differential equations and its application to the radial diffusion system.

Radiation Belt Particle Injection

From figures 6.5a to 6.5c, we see that the posterior probabilities for b, β, and α approximately peak at 0.64, 4.8, and -5 respectively. In figure 6.6a, we see a negative correlation between α and b similar to what was observed in figure 6.4a.

Although the posterior distributions are not uninformative like the uniform priors, they have significantly greater uncertainty as compared to the results of the synthetic problem. This is because of two principal causes.

1. *Data sparsity*: The radiation belt data is sparsely sampled as compared to the synthetic data.
2. *Forward model inadequacy*: The radial diffusion PDE is a simplified model for the dynamics of the radiation belt. This inadequacy of the forward model can have a strong influence in parameter uncertainties.

6.6 Conclusions

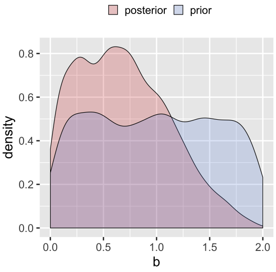
In this chapter, we proposed a surrogate model for the phase space density of particles in the radiation belt, and we presented a method for applying it on the Bayesian inverse problem of quantifying uncertainties in parameters of the simplified radial diffusion model for radiation belt dynamics.

We used our model to perform inference of radial diffusion source term parameters. For this purpose, we tested it on two data sets:

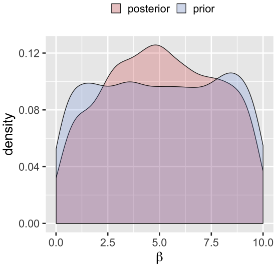
1. synthetically generated phase space density data and
2. in-situ measurements taken by the Van Allen probes.

The model enabled the identification of regions in the parameter space which have a high probability of producing phase space density values which close to the observations.

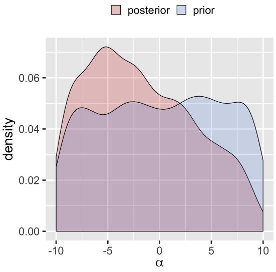
The strength of the proposed method is the ability to quantify the uncertainty in the parameters of a physical system, from a sparse set of observations. Due to the formulation of the surrogate optimisation in its dual form, it allows the inference to scale well with respect to high dimensional basis function expansions. The method can be applied for parameter inference of linear partial differential equations and warrants further research in its improvement.



(a) Parameter b

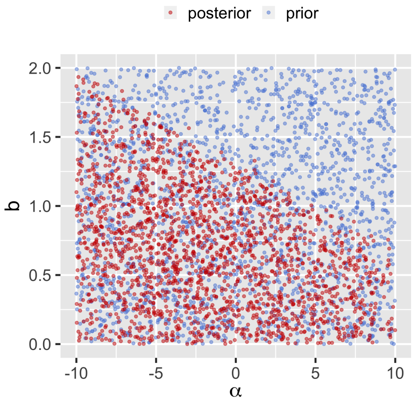


(b) Parameter β

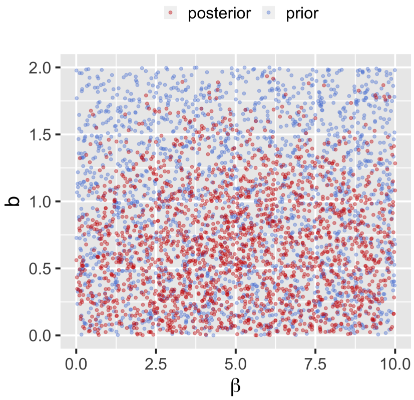


(c) Parameter α

Figure 6.5: **Van Allen Data**: Comparing prior and posterior densities for the parameters of $q(\ell, t)$



(a) Scatter chart: α versus b.



(b) Scatter chart: β versus b.

Figure 6.6: **Van Allen Data**: Prior and posterior samples drawn from parameters of $q(\ell, t)$.

Chapter 7

Forecasting Near-Earth Solar Wind Speed: The DTLR Model

We model the joint regression problem where one signal drives another signal with an unknown time delay, with the forecasting of the solar wind speed based on the Sun's magnetic flux, as the motivating application. This problem, called *dynamic time lag regression* (DTLR) is formalised using a probabilistic setting, modelling the non-stationary time delay between the causes and the effects on the one hand, and the cause-effect relationship on the other hand. A Bayesian approach is presented to tackle the DTLR problem together with theoretical justifications based on linear stability analysis. The approach is empirically validated with proofs of concept on synthetic problems and real-world application of near-Earth solar wind speed prediction.

◆

This chapter is based on research which is under review. Research led by M. Chandorkar, theoretical modeling led by C. Furtlehner, coronal field extrapolations and solar magnetism expertise provided by B. Poduval. M. Sebag and E. Camporeale contributed in supervisory roles.

7.1 Introduction

A significant body of work in machine learning concerns the modeling of spatiotemporal phenomena [Shi and Yeung, 2018, Rangapuram et al., 2018], ranging from markets [Pennacchioli et al., 2014] to weather forecasting [Grover et al., 2015] and space weather prediction [Camporeale et al., 2018a,b, Camporeale, 2019]. This work focuses on the problem of modeling the temporal dependency between two time series, where the latter one is *caused* by the former one [Granger, 1969] with a non-stationary time delay.

7.1.1 Motivation: Forecasting Near-Earth Solar Wind Speed

The Sun, a perennial source of charged energetic particles, drives all geomagnetic phenomena within the Sun-earth system. Specifically, the Sun ejects charged particles into the surrounding space in all directions (solar wind). High speed solar wind is a major threat for the modern world, causing severe damage to satellites, telecommunication infrastructure, under sea pipelines, among others[1]. Interested readers can refer to section 3.2.2 for some historical background to the modern models of the solar wind and the structure Heliospheric Magnetic Field (HMF).

Forecasting near-Earth solar wind speed measured at the L_1 point (see section 3.2.2), based on near-Sun data is a problem of particular importance in space weather prediction due to its large lead time [Munteanu et al., 2013, Haaland et al., 2010]. The challenge of ambient solar wind prediction is two fold. Firstly, although the coronal magnetic field determines the outflow of the solar wind, there are no direct measurements of the coronal magnetic field strength. Secondly the propagation of the solar wind through the interplanetary medium introduces a non-stationary time delay which currently cannot be directly measured.

7.1.2 State Of The Art

For the sake of continuity, we give a quick recap of the state of the art in solar wind forecasting. The reader mostly interested in the machine

---

[1] The adverse impact of space weather is estimated to cost 200 to 400 million per year, but can sporadically lead to much larger losses.

learning method that we develop can skip this section and continue with section 7.1.3.

Research in solar wind forecasting has generally divided the problem into the following components.

1. Using a coronal magnetic field model to extrapolate line of sight photospheric magnetic field measurements, giving an estimation of the coronal magnetic field topology and solar wind flow.

2. Propagation of the coronal solar wind to 1 au (1 au is approximately the distance between the Sun and the Earth).

Reiss et al. [2019] provide an in-depth survey of the state of the art in solar wind prediction, they survey the important coronal magnetic field extrapolation models as well as solar wind propagation procedures. We provide a quick recap for continuity.

The most commonly used coronal magnetic field extrapolation technique is the Potential Field Source Surface model (PFSS) [Altschuler and Newkirk, 1969, Schatten et al., 1969]. The PFSS model assumes a current free (potential) magnetic field structure above the photosphere and expresses the magnetic field $\mathbf{B}$ as the gradient of a scalar magnetic potential $\mathbf{B} = -\nabla\Psi$ which can be solved by constraining the magnetic field to be divergence free ($\nabla^2\Psi = 0$). Since potential fields give closed magnetic fields, a spherical source surface, where the magnetic field is assumed to be radially outwards, is kept as an outer boundary condition. The radius of the spherical source surface is generally set to a height of $2.5R_\odot$, where $R_\odot = 6.957 \times 10^5$ km is the solar radius. The effects of currents have been incorporated in PFSS variants such as the *Potential-Field Current Sheet* (PFCS) [Schatten, 1971] and *Current-Sheet Source Surface* (CSSS) [Zhao and Hoeksema, 1995] models.

It is possible to compute from the solutions of PFSS like models, not only the coronal source surface magnetic field strength, but also the expansion of the magnetic flux tubes of the HMF; the well known *flux-tube expansion* factor ($\mathbf{f}_S$ or FTE). The Wang-Sheeley (WS) model [Wang and Sheeley, 1990] and the improved Wang-Sheeley-Arge (WSA) model [Arge and Pizzo, 2000, Arge et al., 2004] both derive empirical relationships between $\mathbf{f}_S$ computed from the PFSS technique and the source surface solar wind speed v_S.

After computing the coronal magnetic field topology and the source surface solar wind speed, solar wind streams must be propagated to a distance of around 1 au to estimate near-Earth solar wind speeds. Riley and Lionello [2011] provide a survey of various solar wind propagation models, in order of increasing computational complexity.

The simplest propagation technique, known as the *ballistic mapping*, assumes constant velocity propagation from the upper corona ($30R_\odot$) to the Earth, requiring only a longitudinal shift due to solar rotation. Arge and Pizzo [2000] proposed the Arge-Pizzo kinematic evolution scheme meant to be a middle ground between the ballistic mapping and the more complex Magnetohydrodynamics (MHD) models discussed below.

Riley and Lionello [2011] also propose a solar wind propagation technique known as the 1-D Upwind model which uses the inviscid Burger's equation as a simplified model of solar wind flow. The source surface solar wind speed v_S can be mapped using the 1-D Upwind finite difference scheme to 1 au.

The effect of currents is to distort the coronal magnetic field from a current free topology, in order to account for the complex dynamics of solar wind flow. PFSS solutions are often used as boundary conditions for MHD based simulations of the inner heliosphere ($20-30R_\odot$ to 1 au). Common MHD based models include *Magnetohydrodynamics Around a Sphere* (MAS) [Linker et al., 1999], ENLIL [Odstrčil et al., 1996, Odstrčil and Pizzo, 1999a,b, Odstrčil, 2003, Odstrčil et al., 2004] and EUHFORIA [Pomoell and Poedts, 2018].

The most prominent operational solar wind forecasting technique is the hybrid WSA-ENLIL model[2] which consists of the Wang-Sheeley-Arge (WSA) coronal model coupled with the global heliospheric ENLIL model.

Wintoft and Lundstedt [1997] used coronal magnetic field solutions computed by the PFSS model to train a *radial basis function* (RBF) network for predicting the average daily solar wind speed 3 days ahead.

Owens and Riley [2017] used the solutions of MAS model simulated until $30R_\odot$ to construct an ensemble of near-Sun solar wind conditions and forward propagated these conditions to 1 au to give probabilistic forecasts of the near-Earth solar wind speed. Lang and Owens [2019] proposed a variational data assimilation scheme which used the 1-D Upwind model

---

[2] `https://www.swpc.noaa.gov/products/wsa-enlil-solar-wind-prediction`

and solar wind speed measurements from L_1 to improve inner boundary conditions at $30R_{\odot}$.

The current crop of solar wind propagation techniques provide several options for modelers, but they pose one or two key issues: 1. they are computationally intensive 2. they fail to assimilate and learn from data. In this chapter we propose a novel machine learning technique for forecasting near-Earth solar wind speed from the source surface radial magnetic field strength and $\mathbf{f}_S$ computed by the CSSS model as well as the sunspot number and the F10.7 radio flux. Our proposed model works by constructing a probability distribution over possible time delays between near-Sun quantities and near-Earth solar wind observations and then uses the aforementioned probability distribution to formulate a weighted regression problem. The forecasts are assumed to be the output of a neural network architecture. Below we setup the background of our method in the machine learning context.

7.1.3 Predicting What & When

Formally the goal is to model the dependency between heliospheric observations, referred to as *cause series*, and the solar wind speed series recorded at L_1, referred to as *effect series*. The key difficulty is that the time lag between an input and its effect, the solar wind speed recorded at L_1, varies from about 2 to 5 days depending on, among many factors, the initial speed of the solar wind and its interplay with the HMF. Would the lag be constant, the solar wind prediction problem would boil down to a mainstream regression problem. The challenge here is to predict, from solar data $\mathbf{x}(t)$ at time t, the value $y(t+\tau)$ of the solar wind speed reaching the Earth at time $t+\tau$, where the value $y(t+\tau)$ and the time lag τ depend on $\mathbf{x}(t)$.

To our knowledge the regression problem of predicting both *what* the effect is and *when* the effect is observed constitutes a new machine learning problem, that we called *Dynamic Time-Lag Regression* (DTLR). Indeed, the modeling of dependencies among financial time series has been intensively tackled [Zhou and Sornette, 2006]. When considering varying time lag, many approaches rely on dynamic time warping (DTW) [Sakoe and Chiba, 1978]. For instance, DTW is used in Gaskell et al. [2015], taking a Bayesian approach to achieve the temporal alignment of both series under some restricting assumptions (considering slowly varying time lags and linear relationships between the cause and effect time series). More generally, the use of DTW in time series analysis relies on simplifying assumptions on the cause and effect series (same dimensionality and structure) and build upon available cost matrices for the temporal alignment.

This study focuses on the DTLR regression problem and the identification of varying time-lag phenomena involving stochastic dependencies of arbitrary complexity. The originality of the proposed approach compared to the state of the art in DTW series alignment is threefold. Firstly, the cause and effect series are of different dimensionality. While the effect series is scalar, the cause series can be high-dimensional (e.g. images, vectors, etc). Secondly, the relationship between the cause and the effect series can be non-linear (the *what* model). Thirdly, the time lag phenomenon (the *when* model) can be non-smooth (as opposed to e.g. Zhou and Sornette [2006]).

The Bayesian approach proposed to tackle the DTLR regression problem and the associated learning equations are described in section 7.2, followed by a stability analysis and a proof of consistency (section 7.3). The algorithm is detailed in section 7.4. The experimental setting used to validate the approach is presented in section 7.5, and the proofs of concept of the approach are discussed in section 7.6

Notations Given two time series, the cause series $\mathbf{x}(t)$ ($\mathbf{x}(t) \in \mathcal{X} \subset \mathbb{R}^D$) and the observed effect series $y(t)$, the sought model consists of a mapping $f(.)$ which maps each input pattern $\mathbf{x}(t)$ to an output $y(\phi(t))$, and a mapping $g(.)$ which determines the time delay $\phi(t) - t$ between the input and output patterns:

$$y(\phi(t)) = f[\mathbf{x}(t)] \tag{7.1}$$

$$\phi(t) = t + g[\mathbf{x}(t)] \tag{7.2}$$

with

$$f : \mathcal{X} \to \mathbb{R}, \qquad \text{and} \qquad g : \mathcal{X} \to \mathbb{R}^+,$$

where $t \in \mathbb{R}^+$ represents the continuous temporal domain. The input signal $\mathbf{x}(t)$ is possibly high dimensional and contains the hidden cause to the effect $y(t) \in \mathbb{R}$; $y(t)$ is assumed to be scalar in the remainder of this chapter. The mapping $g(.) \in \mathcal{X} \to \mathbb{R}^+$ represents the time delay between inputs and outputs. Vectors are written using bold fonts.

7.2 Probabilistic Dynamically Delayed Regression

As said, equations (7.1) and (7.2) define a regression problem that differs from standard regression along two lines. Firstly, the time lag $g[\mathbf{x}(t)]$ is non-stationary as it depends on $\mathbf{x}(t)$. Secondly, $g[\mathbf{x}(t)]$ is unknown, i.e. it is not recorded explicitly in the training data.

7.2.1 Assumptions

For the sake of the model identifiability and computational stability, the time warping function $\phi(t) = t + g[\mathbf{x}(t)]$ is assumed to be sufficiently regular w.r.t. t. Formally, $\phi(.)$ is assumed to be continuous.

For some authors [Zhou and Sornette, 2006] the monotonicity of $\phi(.)$ is additionally required and enforced using constraints: $\phi(t_1) \leq \phi(t_2), \forall t_1 \leq t_2$. However, this assumption will not be enforced in the model proposed below.

7.2.2 Probabilistic Dynamic Time-Lag Regression

For practical reasons, cause and effect series are sampled at constant rate. In the following they are noted $\mathbf{x}_t$ and y_t, with t to be understood as a discrete time index. Accordingly, the mapping g now outputs a discrete time lag, where the delay $g(\mathbf{x}_t)$ between cause $\mathbf{x}_t$, and effect $y_{t+g(\mathbf{x}_t)}$, ranges in a finite set of integers $T = \{\Delta t_{min}, \ldots, \Delta t_{max} : 0 \leq \Delta t_{min} < \Delta t_{max}\}$ which is defined using domain knowledge.

The unavoidable error due to the discretisation of the continuous time lag and the uncertainty introduced due to the lack of observations of the mapping g are mitigated by using a probabilistic model. The time lag, now a stochastic quantity, is modelled as the vector $\boldsymbol{\tau} = [\tau_1, \ldots, \tau_{|T|}]$ of binary latent variables, where τ_i indicates whether $\mathbf{x}$ drives y_i ($\tau_i = 1$) or not ($\tau_i = 0$). The assumption that every cause has a single effect is modelled by imposing[3]

$$\sum_{i \in T} \tau_i = 1. \tag{7.3}$$

---

[3] Note however that the cause-effect correspondence might be many-to-one, with an effect depending on several causes.

Let $\mathcal{T}$ denote the set of all binary vectors of length $|T|$, which satisfy equation (7.3). In the DTLR framework, the task is to learn two components:

1. a set of independent predictors $\{\hat{y}_i(\mathbf{x}),\ i \in T\}$, and
2. a conditional probability distribution $\hat{p}(\boldsymbol{\tau}|\mathbf{x})$ on $\mathcal{T}$, estimating the probability of the time delay of the effects of x.

For an input pattern $\mathbf{x}_t$, let $\mathbf{y}_t$ denote the vector of random variables $y_{t+i},\ i \in T$. We express the conditional probability $P\big[\mathbf{y}_t|\mathbf{x}_t = \mathbf{x}\big]$ as a mixture of Gaussians[4] centered on the predictors $\hat{y}_i(\mathbf{x})$,

$$P\big[\mathbf{y}_t|\mathbf{x}_t = \mathbf{x}\big] = \sum_{\boldsymbol{\tau}\in\mathcal{T}} \hat{p}\big(\boldsymbol{\tau}|\mathbf{x}\big) \prod_{i\in T} \mathcal{N}\big(\hat{y}_i(\mathbf{x}), \sigma_i(\boldsymbol{\tau})\big)\ , \tag{7.4}$$

where the mixture weights are defined as $p\big(\boldsymbol{\tau}|\mathbf{x}\big)$. Using the constraint defined in equation (7.3), $p\big(\boldsymbol{\tau}|\mathbf{x}\big)$ can be simplified to a vector $\hat{\mathbf{p}}(\mathbf{x})$ consisting of components $[\hat{p}_i(\mathbf{x}) : i \in T]$, such that $\sum_{i\in T} \hat{p}_i(\mathbf{x}) = 1,\ \forall \mathbf{x}$, where $\hat{p}_i(\mathbf{x})$ stands for the probability that the effect of $\mathbf{x}_t = \mathbf{x}$ will occur with delay i.

For further simplifying the analysis, we assume that the variance $\sigma_i^2(\tau)$ of predictor $\hat{y}_i$ does not depend on $\mathbf{x}$, by setting:

$$\sigma_i(\boldsymbol{\tau})^{-2} = \Big(1 + \sum_j \alpha_{ij}\tau_j\Big)\sigma^{-2},$$

with σ^2 a default variance and $\alpha_{ij} \geq 0$ a matrix of non-negative real parameters. The fact that x can influence y_i through predictor $\hat{y}_i(x)$ even when $\tau_i = 0$ reflects an indirect influence due to the auto-correlation of the y series. This influence comes with a higher variance, enforced by making α_{ij} a decreasing function of $|i - j|$. More generally, a large value of α_{ii} compared to α_{ij} for $i \neq j$ corresponds to a small auto-correlation time of the effect series.

Therefore, for any input x a predictive model that learns according to the DTLR framework must output both $\hat{\mathbf{y}}(\mathbf{x}) \in \mathbb{R}^{|T|}$ and $\hat{\mathbf{p}}(\mathbf{x}) \in \mathbb{R}^{|T|}$, with the aim of maximising the probability $P\big[\mathbf{y}_m|\mathbf{x}_t = \mathbf{x}\big]$ given in equation (7.4). In this chapter, $\hat{\mathbf{y}}(\mathbf{x})$ and $\hat{\mathbf{p}}(\mathbf{x})$ are assumed to be the outputs of a neural network. More details on the architecture chosen for this task are given in section 7.4.

---

[4] In many cases, one can map non-Gaussian data into Gaussian using pre-processing

7.2.3 Learning Criterion

Let us denote the dataset as $\{(\mathbf{x},\mathbf{y})\}_{\text{data}}$, $\theta = (\hat{\mathbf{y}},\hat{\mathbf{p}},\sigma,\alpha)$, and empirical averaging on the data is noted as $\mathbb{E}_{\text{data}}$. A heuristic derivation shows that the log-likelihood based on the model in equation (7.4) is approximately given by

$$\mathcal{L}[\{(\mathbf{x},\mathbf{y})\}_{\text{data}}|\theta] = -|T|\log(\sigma) - \mathbb{E}_{\text{data}}\Big[\sum_{i\in T}\frac{1}{2\sigma^2}\big(y_i-\hat{y}_i(\mathbf{x})\big)^2 - \log\big(Z(\mathbf{x},\mathbf{y}|\theta)\big)\Big] \tag{7.5}$$

with the normalisation constant $Z(\mathbf{x},\mathbf{y}|\theta)$ given as:

$$Z(\mathbf{x},\mathbf{y}|\theta) = \sum_{i\in T}\hat{p}_i(\mathbf{x})\exp\Big(-\frac{1}{2\sigma^2}\sum_{j\in T}\alpha_{ji}\big(y_j-\hat{y}_j(\mathbf{x})\big)^2 + \frac{1}{2}\sum_{j\in T}\log(1+\alpha_{ji})\Big).$$

The heuristic derivation of equation (7.5) is given in appendix C. It makes use of a large-sample approximation (assuming infinitely many observations in small volumes $d\mathbf{x}d\mathbf{y}$); while we do not have any guarantees that it is accurate for realistic sample sizes (and if $\mathcal{X}$ is high-dimensional, it may very well not be); but based on optimising equation (7.5) we get excellent predictive results, so we have not sought to find a more precise approximation of the likelihood.

For notational simplicity, the data index t is omitted in the following. The hyper-parameters σ and matrix α of the model are obtained by optimizing $\mathcal{L}$ which gives, for a fixed $\hat{\mathbf{p}}(\mathbf{x})$:

$$\frac{\sigma^2}{1+\alpha_{ij}} = \frac{\mathbb{E}_{\text{data}}\big[\big(y_i-\hat{y}_i(\mathbf{x})\big)^2 q_j(\mathbf{x},\mathbf{y})\big]}{\mathbb{E}_{\text{data}}\big[q_j(\mathbf{x},\mathbf{y})\big]}, \tag{7.6}$$

with the conditional probability $q_i(\mathbf{x},\mathbf{y}) = P(\tau_i = 1|\mathbf{x},\mathbf{y})$ defined as

$$q_i(\mathbf{x},\mathbf{y}) = \frac{1}{Z(\mathbf{x},\mathbf{y}|\theta)}\hat{p}_i(\mathbf{x})\exp\Big(-\frac{1}{2\sigma^2}\sum_{j\in T}\alpha_{ji}\big(y_j-\hat{y}_j(\mathbf{x})\big)^2+\frac{1}{2}\sum_{j\in T}\log(1+\alpha_{ji})\Big) \tag{7.7}$$

These are implicit equations, since $q_i(\mathbf{x},\mathbf{y})$ depends on σ^2, α_{ij}, $\hat{\mathbf{y}}(\mathbf{x})$, and $\hat{\mathbf{p}}(\mathbf{x})$. The proposed algorithm detailed in section 7.4 implements the saddle point method defined in equations (7.6) and (7.7) alternatively. Predictors $\hat{y}_i$ and mixture weights $\hat{p}_i$ are updated through gradient descent based weight updates. The saddle point probability $q_i(x, \mathbf{y})$ is accordingly updated using equation (7.7). The hyper-parameters σ and α_{ij} are subsequently updated using equation (7.6) based on the current $\hat{y}_i$, $\hat{p}_i$, and q_i. Section 7.4 gives details about the precise implementation of this procedure.

7.3 Theoretical Analysis

The proposed DTLR approach is shown to be consistent and analysed in the simple case where α is a diagonal matrix ($\alpha_{ij} = \alpha\delta_{ij}$).

7.3.1 Loss Function & Optimal Predictor

Let us assume that the hyper-parameters of the model have been identified together with predictors $\hat{y}_i(\mathbf{x})$ and weights $\hat{p}_i(\mathbf{x})$. These are leveraged to achieve the prediction of the effect series. For any given input $\mathbf{x}$, the sought eventual predictor is expressed as $(\hat{y}(\mathbf{x}), \hat{I}(\mathbf{x}))$ where $\hat{I}(\mathbf{x})$ is the predicted time lag and $\hat{y}(\mathbf{x})$ the predicted value. The associated L_2 loss is:

$$\mathcal{L}_2(\hat{y}, \hat{I}) = \mathbb{E}_{\text{data}}\Big[\big(y_{\hat{I}(\mathbf{x})} - \hat{y}(\mathbf{x})\big)^2\Big]. \tag{7.8}$$

Then it follows:

Proposition 7.3.1. *with same notations as in equation (7.4), with $\alpha_{ij} = \alpha\delta_{ij}$, $\alpha > 0$, the optimal composite predictor $(y^\star, I^\star)$ is given by*

$$y^\star(x) = \hat{y}_{I^\star(x)}(\mathbf{x}) \qquad \textit{with} \qquad I^\star(\mathbf{x}) = \underset{i}{\operatorname{argmax}}\big(\hat{p}_i(\mathbf{x})\big),$$

Proof. In appendix E. ■

7.3.2 Linear Stability Analysis

The saddle point equations (7.6) and (7.7) admit among others a degenerate solution $\mathbf{p}_0$ corresponding to $\hat{p}_i(\mathbf{x}) = 1/|T|$, $\alpha_{ij} = 0, \forall i, j \in T$, with $\sigma^2 = \sigma_0^2$. Informally the model converges toward this degenerate trivial solution when there is not enough information to build specialised predictors $\hat{y}_i$.

Let us denote $\Delta y_i^2(\mathbf{x}) = \big(y_i - \hat{y}_i(x)\big)^2$ the square error made by predictor $\hat{y}_i$ for $\mathbf{x}$, and

$$\sigma_0^2 = \frac{1}{|T|}\mathbb{E}_{\text{data}}\Big(\sum_{i\in T} \Delta y_i^2(\mathbf{x})\Big)$$

the MSE over the set of the predictors $\hat{y}_i$, $i \in T$.

Let us investigate the conditions under which the degenerate solution may appear, by computing the Hessian of the log-likelihood and its eigenvalues. Under the simplifying assumption

$$\alpha_{ij} = \alpha\delta_{ij},$$

the model involves $2|T|+2$ parameters: α, $r = \sigma^2/\sigma_0^2$, $\hat{\mathbf{y}}$ and $\hat{\mathbf{p}}$. After the computation of the Hessian (appendix D) the system involves three key statistical quantities, two global:

$$C_1[\mathbf{q}] = \frac{1}{\sigma_0^2}\mathbb{E}_{\text{data}}\Big(\sum_{i\in T} q_i(\mathbf{x},\mathbf{y})\Delta y_i^2(\mathbf{x})\Big), \tag{7.9}$$

$$C_2[\mathbf{q}] = \frac{1}{\sigma_0^4}\mathbb{E}_{\text{data}}\Big[\sum_{i\in T} q_i(x,\mathbf{y})\Big(\Delta y_i^2(\mathbf{x}) - \sum_{j\in T} q_j(\mathbf{x},\mathbf{y})\Delta y_j^2(\mathbf{x})\Big)^2\Big], \tag{7.10}$$

and one local:

$$|\mathbf{u}(\mathbf{x})|^2 = \sum_{i\in T} C_{2+i}[\mathbf{x},\mathbf{q}]^2, \tag{7.11}$$

where $\mathbf{u}(\mathbf{x})$ is a $|T|$-vector of components

$$C_{2+i}[\mathbf{x},\mathbf{q}] = \frac{1}{\sigma_0^2}\mathbb{E}_{\text{data}}\Big[q_i(\mathbf{x},\mathbf{y})\big(\Delta y_i^2(\mathbf{x}) - \sum_{j\in T} q_j(\mathbf{x},\mathbf{y})\Delta y_j^2(\mathbf{x})\big)\Big|\mathbf{x}\Big].$$

C_1 represents the covariance between the latent variables $\{\tau_i\}$ and the normalised predictor errors, up to a constant. $C_1 < 1$ indicates a positive correlation between the latent variables and small errors; the smaller the better. For the degenerate solution, i.e. $\mathbf{q} = \mathbf{q}_0$ uniform, $C_1[\mathbf{q}_0] = 1$ and $C_2[\mathbf{q}_0]$ represents the default variability among the prediction errors. $C_{2+i}[\mathbf{x},\mathbf{q}]$ informally measures the quality of predictor $\hat{y}_i$ relatively to the other ones. More precisely, a negative value of $C_{2+i}[\mathbf{x},\mathbf{q}]$ indicates that $\hat{y}_i$ is doing better than average in the neighborhood of $\mathbf{x}$. At a saddle point the parameters are given by:

$$\frac{\sigma^2}{\sigma_0^2} = \frac{|T| - C_1[\mathbf{q}]}{|T|-1} \qquad \text{and} \qquad \alpha = \frac{|T|}{|T|-1}\frac{1-C_1[\mathbf{q}]}{C_1[\mathbf{q}]}.$$

The predictors $\hat{\mathbf{y}}$ are decoupled from the rest whenever they are centered, which we assume. If $\hat{\mathbf{p}}$ is fixed, a saddle point is stable if and only if

$$C_2[\mathbf{q}] < 2C_1^2[\mathbf{q}] + \mathcal{O}(\frac{1}{|T|}).$$

In particular, the degenerate solution is unstable if

$$C_2[\mathbf{q}_0] > 2\big(1 - \frac{1}{|T|}\big).$$

Note that when $\Delta y_i(\mathbf{x})$ are i.i.d and centered, with variance σ_0^2 and relative kurtosis κ (conditionally to $\mathbf{x}$) one has $C_2 = (2+\kappa)(1-1/|T|)$. Therefore, whenever $\Delta y_i^2(\mathbf{x})$ fluctuates and the relative kurtosis is non-negative, the degenerate solution is unstable and will thus be avoided.

When $\hat{\mathbf{p}}(\mathbf{x})$ is iteratively refined by successive weight updates, the degenerate trivial solution becomes unstable as soon as $|\mathbf{u}(\mathbf{x})|^2$ is non-zero, due to the fact that the gradient is inversely proportional to $\mathbf{u}(\mathbf{x})$ (with $d\hat{\mathbf{p}}(\mathbf{x}) \propto -|\mathbf{u}(\mathbf{x})|^2\mathbf{u}(x)$), thus rewarding the predictors with lowest errors by increasing their weights. The system is then driven toward other solutions, among which the localised solutions of the form:

$$\hat{p}_i(\mathbf{x}) = \delta_{iI(\mathbf{x})},$$

with an input dependent index $I(\mathbf{x}) \in T$. As shown in appendix D, the solution of highest likelihood of this type is also optimal with respect to the loss function (equation (7.8)). The stability of such localised solutions and the existence of other (non-localised) solutions is left for further work.

7.4 Overview Of The DTLR Algorithm

```
Initialization of $\alpha$ and $\sigma$
$it \longleftarrow 0$ ;
while $it < max$ do
    while $epoch$ do
        $\boldsymbol{\theta}_{NN} \longleftarrow Optimise(\mathcal{L}(\boldsymbol{\theta}_{NN}, \alpha, \sigma^2))$ ;
    end
    $\sigma^2 \longleftarrow \sigma_0^2 \frac{|T| - C_1[\mathbf{q}]}{|T|-1}$ ;
    $\alpha \longleftarrow \frac{|T|}{|T|-1} \frac{1 - C_1[\mathbf{q}]}{C_1[\mathbf{q}]}$ ;
end
Result: Network weights $\boldsymbol{\theta}_{NN}$, hyper-parameters $\alpha, \sigma^2$
```

Algorithm 1: DTLR algorithm

The DTLR algorithm learns both regression models $\hat{\mathbf{y}}(\mathbf{x})$ and $\hat{\mathbf{p}}(\mathbf{x})$ from series $\mathbf{x}_t$ and y_t, using alternate optimization of the model parameters and the model hyper-parameters α and σ^2, after equations (7.6) and (7.7). The model search space is that of neural nets, parametrized by their weight vector $\boldsymbol{\theta}_{NN}$. The inner optimization loop updates $\boldsymbol{\theta}_{NN}$ using mini-batch based stochastic gradient descent. At the end of each epoch, after all mini-batches have been considered, the outer optimization loop computes hyper-parameters α and σ^2 on the whole data.

The initialization of hyper-parameters α and σ is settled using preliminary experiments (same setting for all considered problems: $\alpha \sim U(0.75, 2)$; $\sigma^2 \sim U(10^{-5}, 5)$).

The neural architecture implements predictors $\hat{\mathbf{y}}(\mathbf{x})$ and weights $\hat{\mathbf{p}}(\mathbf{x})$ on the top of a same feature extractor from input $\mathbf{x}$. The architecture of the feature extractor is a two-hidden layer fully connected network. On the top of the feature extractor are the one-layer $\hat{\mathbf{y}}$ and $\hat{\mathbf{p}}$ models, each with $|T|$ output neurons, with $|T|$ the size of the chosen interval for the time lag.

Table 7.1: Network Architecture Details

Problem	# Hidden layers	Layer sizes	Activations
I	2	$[40, 40]$	[ReLU, Sigmoid]
II	2	$[40, 40]$	[ReLU, Sigmoid]
III	2	$[40, 40]$	[ReLU, Sigmoid]
IV	2	$[60, 40]$	[ReLU, Sigmoid]
Solar Wind	2	$[50, 50]$	[ReLU, Sigmoid]

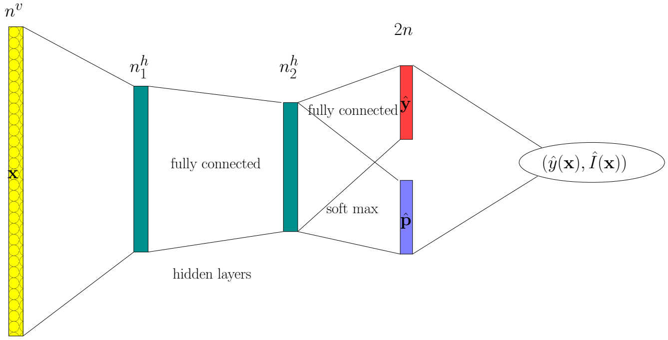


Figure 7.1: Architecture of the neural network specified by the number of units $(n^v, n^h_1, n^h_2, 2|T|)$ in each layer.

7.5 Experimental Setting

The goal of the experiments is twofold. Firstly, the DTLR predictive performance is assessed by considering 1. the RMSE of the predicted effect series $\hat{y}_t$, computed from Eq. (7.3.1) 2. the accuracy of the time lag prediction. The latter performance indicator is measured and compared to the ground truth using synthetic problems, detailed below: although time lag relationships do exist in real world data sets [Munteanu et al., 2013, Zhou and Sornette, 2006], we are not aware of datasets with time lag relationships explicitly annotated. The former performance indicator is comparatively assessed using the naive baseline, the regression model computed by assuming a fixed time lag set to $\frac{\Delta t_{min} + \Delta t_{max}}{2}$. The Pearson correlation of between y_t and the predicted $\hat{y}_t$ series is also considered as overall performance indicator of the prediction.

Table 7.2: Synthetic and Real-World Problems. For the solar wind problem, training and test data sizes represent one cross validation fold

Problem	# train	# test	d	$\|T\|$
I	$10,000$	$2,000$	10	15
II	$10,000$	$2,000$	10	20
III	$10,000$	$2,000$	10	20
IV	$10,000$	$2,000$	10	20
Solar Wind	$77,367$	648	374	12

The second goal of experiments is to determine how informative are the key statistical quantities σ_0 and C_1 (section 7.3.2), and whether they can effectively be used as measures of confidence about the prediction results. Table 7.2 summarises the dimensions of the synthetic and real problems used as proofs of concept for the DTLR validation.

Synthetic Problems. Four synthetic problems of increasing difficulty are generated using *Stochastic Langevin Dynamics.* In all problems, the cause signal $\mathbf{x}_t \in \mathbb{R}^{10}$ and the effect signal y_t are generated as follows (with $\eta = 0.02, s^2 = 0.7$):

$$\mathbf{x}_{t+1} = (1-\eta)\mathbf{x}_t + \mathcal{N}(0, s^2) \tag{7.12}$$

$$v_t = k||\mathbf{x}_t||^2 + c \tag{7.13}$$

$$y_{t+g(\mathbf{x}_t)} = f(v_t), \tag{7.14}$$

with time-lag mapping $g(\mathbf{x}_t)$ ranges in a time interval with width 20 (except for problem I where $|T| = 15$). The complexity of the synthetic problems is governed by the amplitude and time-lag functions f and g, as shown in the table below.

Problem	$f(v_t)$	$g(\mathbf{x}_t)$	Other
I	v_t	5	$k = 10, c = 0$
II	v_t	$100/v_t$	$k = 1, c = 10$
III	$\sqrt{v_t^2 + 2ad}$	$(\sqrt{v_t^2 + 2ad} - v)/a$	$k = 5, a = 5, d = 1000, c = 100$
IV	v_t	$g(\mathbf{x}_t) = \exp(v_t) / (1 + \exp(v_t/20))$	$k = 10, c = 40$

Solar Wind Speed Prediction As said in section 7.1.1, the task of predicting solar wind speed from heliospheric data not only has scientific significance; it is also challenging due to the distance between the Sun and the Earth and the non-stationary propagation time of the solar plasma through the interplanetary medium.

The inputs $\mathbf{x}_t$ are compiled from two sources, synoptic Carrington maps of the photospheric magnetic field taken from the *Global Oscillation Network Group* (GONG) and solar activity proxies taken from the OMNI data set[5]. Specifically, $\mathbf{x}_t$ is a vector consisting of the components outlined in table 7.3.

For each input pattern, time lagged solar wind data is extracted corresponding to minimum and maximum time delays of two and five days respectively. For computational convenience, each three day time window is pre-processed by computing sliding six hour medians yielding $|T| = 12$ time slots[6]. Before training, the solar wind data was mapped into standardized Gaussian space by applying a quantile-quantile followed by the inverse probit mapping.

DTLR is validated using a 9 fold cross-validation, where the test data consists of one (continuous) Carrington rotation (see table 7.4). The performance on Carrington rotation 2077 (first fold in table 7.4) is compared with the state of the art [Riley and Lionello, 2011] in table 7.6, while the overall cross-validation performance is compared with a fixed time lag baseline which uses the same inputs as the DTLR model (table 7.3) and has the same architecture until the penultimate layer (figure 7.1). For Carrington rotations 2077 and 2184, the hourly solar wind time series reconstructions are shown in figures 7.6a and 7.6b respectively.

---

[5] `https://omniweb.gsfc.nasa.gov`

[6] Before computing the cross-validation performance, the predictions are mapped back to hourly resolution using interpolation

Quantity	**Source**	**Domain**	**Notes**
$\log \mathbf{f}_S$	GONG	$\mathbb{R}^{180}$	$\mathbf{f}_S$ or FTE is computed from the outputs of the CSSS model.
$\mathbf{B}_{cp}$	GONG	$\mathbb{R}^{180}$	The radial magnetic field strength on the solar cusp surface is the primary output computed by the CSSS model.
$\mathbf{v}_{27}$	OMNI	$\mathbb{R}^{\lvert T \rvert}$	$\mathbf{v}_{27}$ is the solar wind speed recorded 27 days prior, one for each value of the forward time window.
SSN	OMNI	$\mathbb{R}^+$	The sun spot number measures the number of visible sun spots on the solar disk.
F10.7	OMNI	$\mathbb{R}^+$	F10.7 is the measured solar radio flux.

Table 7.3: Inputs used in the DTLR solar wind forecast model.

Table 7.4: Cross validation splits used to evaluate DTLR on the solar wind forecasting task

Split Id	**Carrington Rotation**	**Start**	**End**
1	2077	2008/11/20 07:00:04	2008/12/17 14:38:34
2	2090	2009/11/09 20:33:43	2009/12/07 04:03:59
3	2104	2010/11/26 17:32:44	2010/12/24 01:15:56
4	2117	2011/11/16 07:04:41	2011/12/13 14:39:28
5	2130	2012/11/04 20:39:43	2012/12/02 04:06:23
6	2143	2013/10/25 10:17:52	2013/11/21 17:36:35
7	2157	2014/11/11 07:09:56	2014/12/08 14:41:02
8	2171	2015/11/28 04:09:27	2015/12/25 11:53:33
9	2184	2016/11/16 17:41:04	2016/12/14 01:16:43

7.6 Empirical validation

Table 7.5 summarises the DTLR performance on the synthetic and real-world problems, respectively compared to the naive baseline (constant time lag) and to the state of the art for the real-world solar wind problem.

The values of the σ_0 and C_1 quantities involved in the stability analysis (section 7.3.2) are also reported. As said, $C_1 < 1$ indicates a specialization among predictors found by the solution. The comparison of σ_0 and the RMSE indicates how better the learned model is compared to the trivial degenerate solution (uniform $\hat{\mathbf{p}}$, assigning an equal weight to all $\hat{y}_i$). Finally, the Pearson correlation between $\hat{y}_t$ and y_t is reported; while its absolute value is less informative than it appears due to the auto-correlation of the series, it allows to compare different predictors.

On the easy Problem I, the model predicts the correct time lag for 97.93% of the samples. The higher value of σ_0 in problems I and II compared to the other problems is explained from the higher variance in the generated time series $y(t)$.

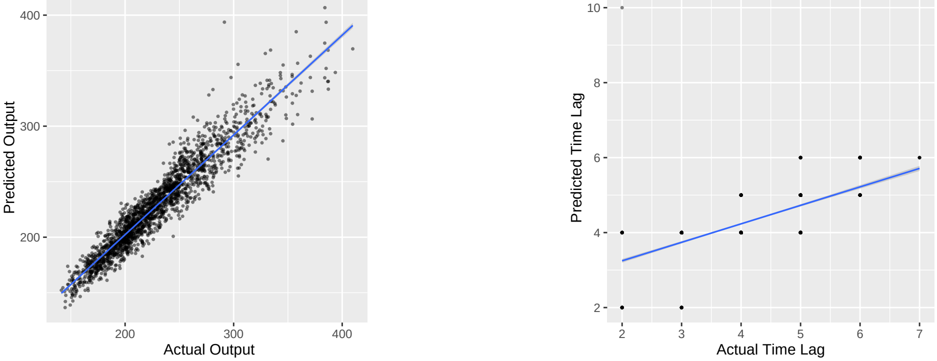
On Problem II, the model accurately learns the inverse relationship between $\mathbf{x}_t$, $g(\mathbf{x}_t)$ and y_t on average. The time lag is overestimated in the regions with low time lag (with high velocity), which is blamed on the low sample density in this region, due to the data generation process.

Interestingly, Problems III and IV are better handled by DTLR, despite a more complex dynamic time lag relationship. In both latter cases however, the model tends to under-estimate the time lag in the high time lag regions and conversely to over-estimate it in the low time lag region.

Concerning the solar wind problem, DTLR shows encouraging results on the cross-validation experiments as can be seen in table 7.5 and visualised in figure 7.5. The significantly higher difficulty of the solar wind forecasting

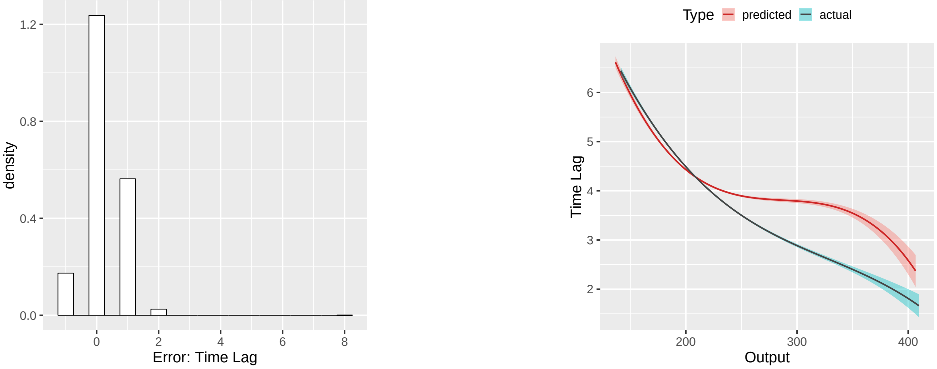
Table 7.5: Performance: DTLR / Base Line / DTLR Time Lag Prediction

Problem	M.A.E	R.M.S.E	Pearson Corr.	σ_0	C_1
Pb I	8.82 / 21.79 / 0.021	12.35 / 28.79 / 0.26	0.98 / 0.87 / –	29.8	0.14
Pb II	10.15 / 27.40 / 0.4	13.70 / 35.11 / 0.67	0.95 / 0.73 / 0.70	26.83	0.16
Pb III	3.17 / 11.01 / 0.17	4.63 / 14.99 / 0.42	0.98 / 0.79 / 0.84	11.84	0.09
Pb IV	3.88 / 12.28 / 0.34	5.33 / 15.89 / 0.64	0.98 /0.79/ 0.81	12.18	0.13
Solar Wind	56.35 / 66.45 / –	74.20 / 84.53 / –	0.6 / 0.41 / –	76.46	0.89



(a) **Problem II**, Goodness of fit, Output $y(x)$

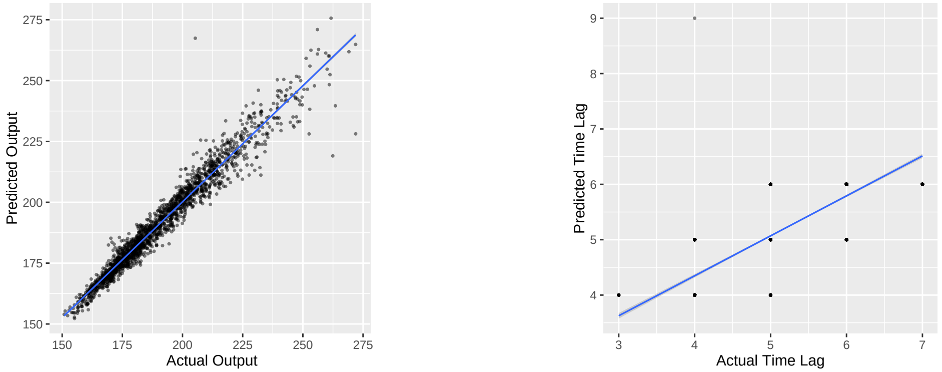
(b) **Problem II**, Goodness of fit, Time lag $\tau(t)$



(c) **Problem II**, Error of time lag prediction

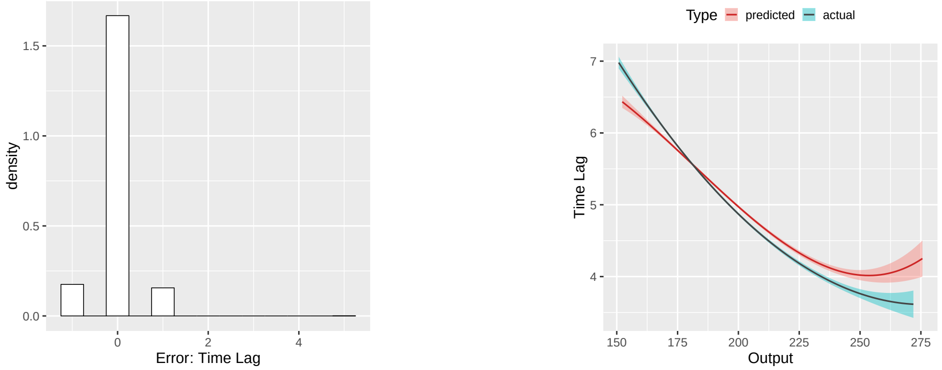
(d) **Problem II**, Output vs Time Lag Relationship

Figure 7.2: **Problem II**, Results



(a) **Problem III**, Goodness of fit, Output $y(x)$

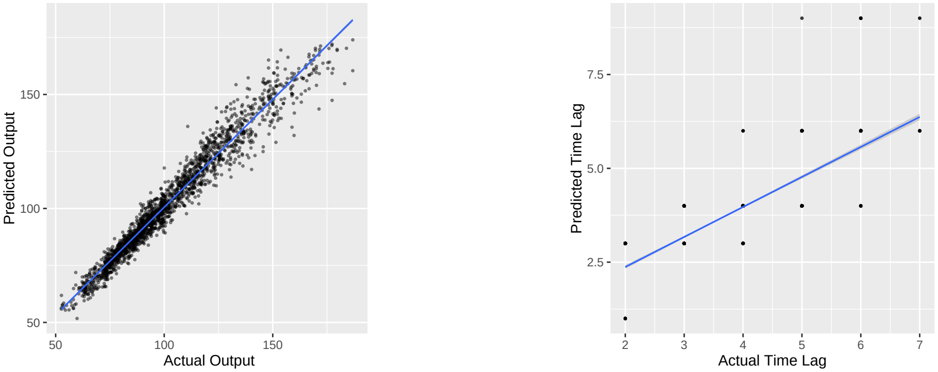
(b) **Problem III**, Goodness of fit, Time lag $\tau(t)$



(c) **Problem III**, Error of time lag prediction

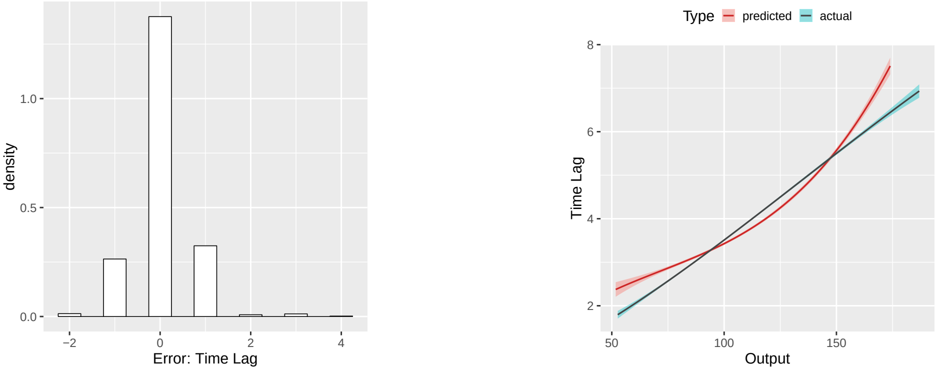
(d) **Problem III**, Output vs Time Lag Relationship

Figure 7.3: **Problem III**, Results



(a) **Problem IV**, Goodness of fit, Output $y(x)$

(b) **Problem IV**, Goodness of fit, Time lag $\tau(t)$



(c) **Problem IV**, Error of time lag prediction

(d) **Problem IV**, Output vs Time Lag Relationship

Figure 7.4: **Problem IV**, Results

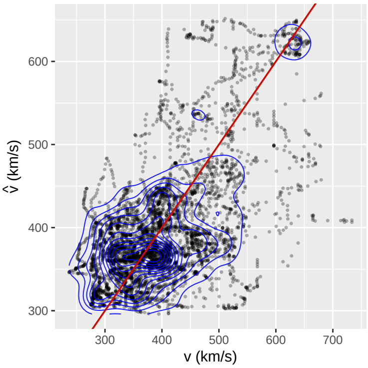
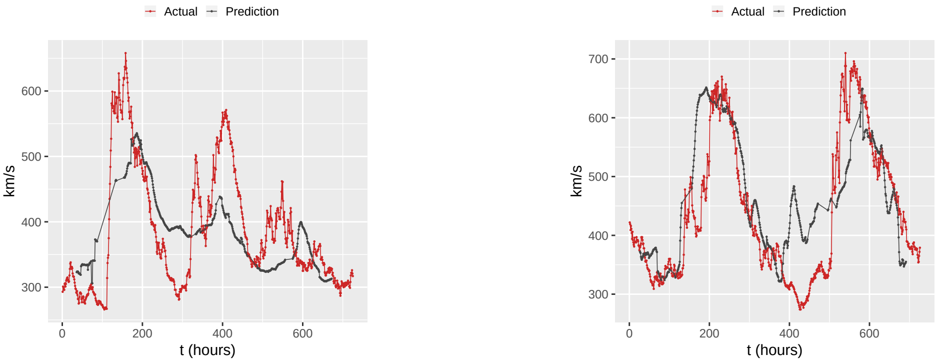


Figure 7.5: Predicted vs Actual Solar Wind Speed, red diagonal represents a perfect prediction and blue lines are contours.

problem is witnessed by the C_1 value close to the degenerate value of 1. In table 7.6, we compare the performance of the DTLR model and the fixed time lag baseline with the state of the art from Reiss et al. [2019], on the solar wind data from Carrington rotation 2077 (see table 7.4). The DTLR model gives improvements in predictive performance.

Table 7.6: Performance Comparison on CR 2077: DTLR , Fixed Lag Base Line vs Reiss et al. [2019]

Model	M.A.E	R.M.S.E
WS	74.09	85.27
DCHB	83.83	103.43
WSA	68.54	82.62
Ensemble Median (WS)	71.52	83.36
Ensemble Median (DCHB)	78.27	100.04
Ensemble Median (WSA)	62.24	74.86
Persistence (4 days)	130.48	161.99
Persistence (27 days)	66.54	78.86
Fixed Lag Baseline	67.33	80.39
DTLR	60.19	72.64



(a) Hourly forecasts for the period 2008-11-20 07:00 to 2008-12-17 14:00

(b) Hourly forecasts for the period 2016-11-16 17:00 to 2016-12-14 01:00

Figure 7.6: **Solar Wind Prediction**: reconstructed time series predictions

7.7 Conclusions

The contribution of the work is twofold. Firstly, we define a new ML setting, motivated by an important scientific and practical problem from the domain of space weather, emphasizing that this real-world problem is open for over two decades. This ML setting, called Dynamic Time Lag Regression, is concerned with the inference of lagged causal relationships between time series.

Secondly, the proposed DTLR formalization supports the definition of a nested inference procedure, relying on a saddle point optimization process. A closed form analysis of the stability of the inferred model under simplifying assumptions has been conducted, yielding a practical alternate optimization formulation, implemented in the DTLR algorithm. The approach demonstrates its merits with some proofs of concept on synthetic problems considering time lag models with diverse complexity. The application on our motivating real-world problem shows the potential of the approach, considering that the DTLR model involves no domain knowledge in the pre-processing of the data or in the sought prediction model. From an applicative perspective, a next step toward improving the predictive performances will consist of enriching the data sources and the description of the cause series $\mathbf{x}_t$.

On the methodological side, the longer term research perspective consists of extending the proposed nested inference procedure and integrating the model selection step within the inference architecture; the challenge is to provide the algorithm with the means of assessing online the stability and/or the degeneracy of the learning trajectory.

Chapter 8

Concluding Remarks

This thesis represents an exploration of the possibilities for using machine learning techniques for advancing space weather research. The work presented here was classified into three principal research problems or themes as mentioned in chapter 1. Below we give a quick summary of the main achievements of this thesis and avenues for further research.

8.1 Discussion

The guiding motivation of this thesis was the creation of next-generation space weather forecasting and monitoring systems. From an operational perspective, this goal is a challenging and long-term endeavour. From the point of view of research, the work presented in this thesis breaks down the Sun-Earth system into components and establishes important paradigms for building forecasting and inference systems for each component. Below, we give a more detailed overview of the achievements of this work and some further research questions.

Geomagnetic Time Series Forecasting

Using Gaussian process auto-regressive methods, it is possible to obtain accurate and reliable probabilistic forecasts for the Dst index, up to five hours ahead. By viewing the prediction problem from a probabilistic lens, the GP-AR, GP-ARX, and GPNN models proposed in chapters 4 and 5 represent an important paradigm shift for geomagnetic activity modeling. The ability to provide probabilistic forecasts will become an indispensable component of future space weather monitoring systems.

The GP-AR and GP-ARX models give a general framework for modeling non-linear dynamical systems and provide uncertainty estimates on their time evolution. Their main drawback is the $O(N^3)$ time complexity for performing inference which makes applications on large data sets challenging. By using neural network based models as mean functions of GP models, one can create hybrid models which somewhat circumvent this drawback while retaining the probabilistic forecasting capabilities.

Radiation Belt Parameter Inference

In chapter 6, we proposed a machine learning based surrogate model which is informed by models of physical dynamics and can compute likelihoods over irregularly spaced observations.

Although physics informed machine learning models may not offer benefits over finite difference, finite volume, and finite element based methods for solving forward problems, when solving inverse problems they have some desirable properties: the ability to synthesise irregular observations and physics models into a unified methodology for parameter inference and uncertainty quantification.

When performing inference over the parameters of the radiation belt dynamics, one needs to take into account the sensitivity of the radiation belt model to its parameters. It is advisable to use domain knowledge and sensitivity analysis to constrain the numerical ranges of the parameter prior distributions as this aides identification and obtaining compact uncertainty estimates.

Casting the surrogate optimisation problem (equation (6.5)) in its dual form enables the use of potentially infinite dimensional basis function expansions but it introduces the same computational challenges that come with Gaussian process inference.

Physics informed machine learning models have the potential to play an important role in data assimilation and probabilistic inference, the uncertainties obtained for unobserved physical parameters can provide qualitative indications about the adequacy of a particular physical model.

Solar Wind Prediction

The effect of time lag relationships between interacting systems can impact the performance of predictive models which give forecasts for a fixed time horizon. In chapter 7, we proposed the dynamic time-lag regression (DTLR) framework, a novel machine learning approach for training predictive models on time series data sets which have non-stationary time lag dependencies.

The task of predicting near-Earth solar wind speed from heliospheric data is very challenging and requires astute application of all the tools at our disposal: 1. models of the heliospheric magnetic field, 2. machine learning techniques, and 3. solar and near-Earth data. Using the DTLR framework, we were able to achieve encouraging progress in the forecasting of near-Earth solar wind speed.

8.2 Further Research

Although the space weather problem is the central motivation for the work presented here, the techniques presented in this thesis are generally applicable in the modeling and forecasting of physical systems. Some possible questions for further research are listed below.

1. With the existing state of the art, how far can we extend the time horizon of geomagnetic activity forecasts? What kind of data and methods will be important in making ten-hour-ahead or twelve-hour-ahead forecasts of the Dst index?

2. How do we extend the phase space density surrogate model to higher dimensional radiation belt dynamics, i.e. diffusion across all three adiabatic invariants? How do we deal with the computational challenges that arise from performing inference over parameters of 3-d diffusion models and large scale data sets?

3. When working in the context of PDE constrained inverse problems, how do we separate uncertainties arising from parameter identifiability and forward model inadequacy? How do we perform inference over the parameters of a non-linear PDE using machine learning based surrogate models?

4. How can we improve the accuracy of solar wind forecasts made by the DTLR model? Can we combine the current sheet source surface (CSSS) model and DTLR models into an operational solar wind forecasting system? Is it beneficial to use a surrogate in place of the CSSS model to compute the topology of the coronal magnetic field?

5. What are the possible applications of the DTLR model in domains other than space weather? Is the DTLR model suitable for forecasting applications in finance or energy grids?

Appendices

Appendix A

Notes On Gaussian Process Time Series Models

Gaussian processes provide a systematic and flexible framework for probabilistic inference in machine learning. Their formulation is general and their existence is dependent on two key conditions outlined in section 4.2 which we restate here.

For any input space $\mathcal{X}$, a real valued scalar GP (equation (A.1)) can be created given two functions.

1. A mean function $m : \mathcal{X} \longrightarrow \mathbb{R}$.

2. A symmetric positive definite covariance function (kernel) $K : \mathcal{X} \times \mathcal{X} \longrightarrow \mathbb{R}^+$ [Berlinet and Thomas-Agnan, 2004, ch. 1&2].

$$f(x) \sim \mathcal{GP}(m(x), K(x, x')) \tag{A.1}$$

It is important to note that there are no restrictions placed on the input space $\mathcal{X}$ in GPs. The inputs can be continuous, discrete or have complex structure. Rasmussen and Williams [2005, ch. 4, sec. 4.4] give some examples of covariance functions defined over the space of strings (sequences of characters drawn from a finite alphabet) and note that it is indeed possible to construct covariance functions over structured objects such as trees and general graphs.

Gaussian process time series modelling has a rich body of work [Turner, 2012, Frigola-Alcalde, 2016]. There are two ways in which GP models can be applied to time series data: 1. explicit time & 2. implicit time

Explicit vs Implicit Representation

Depending on whether or not time appears directly in the inputs space of a GP, we can classify a GP time series model as explicit or implicit respectively. Notationally, explicit and implicit time GP formulations are shown in equation (A.2) and equation (A.3) respectively.

$$y(t) \sim \mathcal{GP}(m(t), K(t, s)) \tag{A.2}$$

$$y(t) \sim \mathcal{GP}(m(\mathbf{x}(t)), K(\mathbf{x}(t), \mathbf{x}(s))) \tag{A.3}$$

In the explicit time GP model, the mean and covariance of the process are functions of the continuous time coordinate t. On the other hand in implicit time GP models, the mean and covariance are functions of some state space $\mathbf{x}_t$. This key difference has significant implications for the system trajectories and uncertainty characteristics.

In section 4.3, we formulated the Dst prediction model according to equation (A.3). By setting $\mathbf{x}(t)$ to a time history of Dst and possibly solar wind parameters, we built the GP-AR and GP-ARX prediction models.

GP-AR & GP-ARX Models

Implicit time GP models (equation (A.3)) define probabilistic dynamics for a system $y(t)$ in terms of some time varying system state $\mathbf{x}(t)$. To define their finite dimensional distributions using the GP methodology, we discretely subsample $y(t)$ and (t) and denote their discrete analogues as y_t and $\mathbf{x}_t$ respectively.

We give a detailed picture of the GP-AR model, its finite dimensional distribution, its sampling and dynamics. The GP-ARX model is a straight forward extension of the GP-AR dynamics shown below.

We first setup some background notation. The GP-AR system of order p, is determined by the system state $\mathbf{x}_t$; a dimensional vector composed of the time history of the system, shown in equation (A.4).

$$\mathbf{x}_t = \begin{bmatrix} y_{t-1} \\ \vdots \\ y_{t-p} \end{bmatrix}_{p \times 1} \tag{A.4}$$

For every time step t, we combine all system states $\mathbf{x}_t, \cdots, \mathbf{x}_{p-1}$ into a design matrix $\mathbf{X}_t$ shown in equation (A.5). The system trajectory $\mathbf{y}_t$ in equation (A.6) is the path taken by the system from time step p to time step t.

$$\mathbf{X}_t = \begin{bmatrix} \mathbf{x}_t^T \\ \vdots \\ \mathbf{x}_p^T \end{bmatrix}_{(t-p+1) \times p} \tag{A.5}$$

$$\mathbf{y}_t = \begin{bmatrix} y_t \\ \vdots \\ y_p \end{bmatrix}_{(t-p+1) \times 1} \tag{A.6}$$

Joint Distribution

The joint distribution of the system trajectory conditional on its initial state is a multivariate Gaussian as shown in equation (A.7). The mean and covariance of the joint distribution are calculated using the mean function $m(.)$ and kernel $K(.,.)$ from equation (A.3).

Although the joint distribution of the system trajectory is straight forward to compute, sampling trajectories from the joint distribution is intractable because elements of the mean vector and covariance matrix are not computable unless the entire trajectory is known beforehand.

To circumvent the sampling intractability of equation (A.7), we use a probabilistic simulation based approach to the GP-AR process.

$$y_p, \cdots, y_t | y_0, \cdots, y_{p-1} \sim \mathcal{N}\left(\begin{bmatrix} m(\mathbf{x}_p) \\ \vdots \\ m(\mathbf{x}_t) \end{bmatrix}, \begin{bmatrix} K(\mathbf{x}_p, \mathbf{x}_p) & \cdots & K(\mathbf{x}_p, \mathbf{x}_t) \\ \vdots & \ddots & \vdots \\ K(\mathbf{x}_t, \mathbf{x}_p) & \cdots & K(\mathbf{x}_t, \mathbf{x}_t) \end{bmatrix} \right) \tag{A.7}$$

GP-AR: Dynamics & Sampling

GP-AR and GP-ARX models can be understood as probabilistic simulation models. The iterative dynamics of GP-AR is captured in equations (A.8) and (A.9). The initial seed of the process is the system trajectory from time step 0 to time step $p-1$ which is sampled from a standard multivariate Gaussian distribution. At each time step t, y_t conditional to the system trajectory $\mathbf{y}_{t-1}$ is sampled from a Gaussian distribution (equation (A.9)) whose mean and covariance are calculated by appropriately conditioning the joint distribution in equation (A.7) and constructing the system state $\mathbf{x}_t$ and design matrix $\mathbf{X}_{t-1}$.

$$(y_0, \cdots, y_{p-1}) \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{A.8}$$

$$y_t | \mathbf{y}_{t-1} \sim \mathcal{N}(\bar{m}_t, \bar{\sigma}_t^2) \tag{A.9}$$

$$\begin{aligned} \bar{m}_t &= m(\mathbf{x}_t) + \mathbf{k}_t^T \mathbf{K}_t^{-1} \left(\mathbf{y}_t - m(\mathbf{X}_{t-1})\right) \\ \bar{\sigma}_t^2 &= K(\mathbf{x}_t, \mathbf{x}_t) - \mathbf{k}_t^T \mathbf{K}_t^{-1} \mathbf{k}_t \end{aligned} \tag{A.10}$$

In equation (A.10), the mean and variance of the conditional distribution of $y_t|\mathbf{y}_{t-1}$ are computed, where $\mathbf{K}_t = [K(\mathbf{X}_{t-1}, \mathbf{X}_{t-1})]_{(t-p)\times(t-p)}$ is the covariance matrix computed between each pair of rows of $\mathbf{X}_{t-1}$ and $\mathbf{k}_t = [K(\mathbf{x}_t, \mathbf{X}_{t-1})]_{1\times(t-p)}$ is the cross covariance matrix computed between the input features $\mathbf{x}_t$ and each row of the $\mathbf{X}_{t-1}$.

Relationship with Time Series Models : AR(p)

Due to the abstract nature of GPs, they generalize common time series models. For example, consider AR(p), the family of discrete auto-regressive time series models shown in equations (A.11) and (A.12). At each time step, the systems state y_t is determined by a linear combination of p previous system states $y_{t-1}, \cdots, y_{t-p}$ plus Gaussian noise.

$$y_t = \sum_{k=1}^{p} \beta_k y_{t-k} + z_t \tag{A.11}$$

$$z_t \sim \mathcal{N}(0, \sigma_\varepsilon^2) \tag{A.12}$$

We can write equations (A.11) and (A.12) in probabilistic form as:

$$y_t|y_{t-1}, \cdots, y_{t-p} \sim \mathcal{N}(\sum_{k=1}^{p} \beta_k y_{t-k}, \sigma_\varepsilon^2)$$

We can see that the GP-AR model family contains the AR(p) model in equation (A.11) as a special case. By setting $m(\mathbf{x}_t) = \sum_{k=1}^{p} \beta_k Y_{t-k}$ and $K(\mathbf{x}_t, \mathbf{x}_s) = \sigma_\varepsilon^2 \delta(\mathbf{x}_t, \mathbf{x}_s)$, where $\delta(.,.)$ is the Dirac delta function, we obtain the conditional probability density of the random variable $y_t|y_{t-1}, \cdots, y_{t-p}$ described above. Roberts et al. [2013] note that the AR(p) process is also equivalent to an explicit time GP model having $m(t) = \mathbb{E}[y(0)]$ and a Matérn covariance function

$$K(t, s) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu} \frac{|t-s|}{\rho} \right)^{\nu} \mathcal{K}_\nu \left(\sqrt{2\nu} \frac{|t-s|}{\rho} \right)$$

with $\nu = p + \frac{1}{2}$, where $\Gamma(.)$ the gamma function, $\mathcal{K}_\nu(.)$ the modified Bessel function of the second kind, ρ being the length scale and σ^2 the amplitude of $K(.,.)$.

The advantage of the implicit time GP-AR model over the explicit time GP model is that the GP-AR family can also simulate non-linear auto-regressive processes. This is because the kernel creates a Hilbert space $\mathcal{H}(\mathbb{C})$ spanned by a basis of non-linear orthogonal basis functions $\phi \in \mathcal{H}(\mathbb{C})$, such that the kernel can be decomposed [Rasmussen and Williams, 2005, sec. 4.3] in the following manner.

$$K(\mathbf{x}_t, \mathbf{x}_s) = \sum_{i=1}^{\infty} \lambda_i \phi(\mathbf{x}_t) \phi(\bar{\mathbf{x}}_s)$$

Rasmussen and Williams [2005, ch. 2] also notes that 'Gaussian process regression can be viewed as Bayesian linear regression with a possibly infinite number of basis functions'. We provide a simple example of how the GP-AR model can approximate the dynamics of a non-linear auto-regressive system.

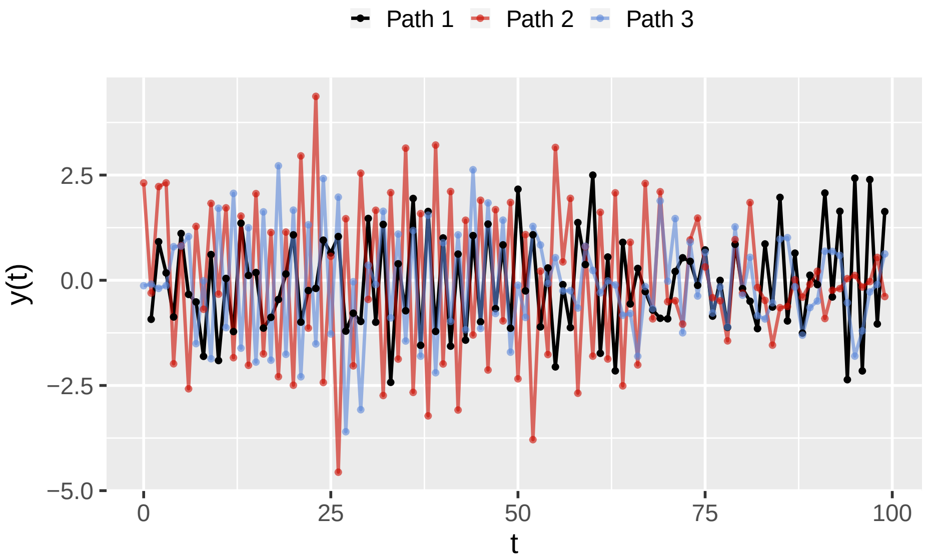


Figure A.1: Samples drawn from non-linear autoregressive model in equation (A.13). The path coloured in black is used partially ($t \leq 40$) in the training and the rest as test data

Example: System Identification

Consider y_t, a quadratic auto-regressive system of order three, expressed below.

$$y_0 \sim \mathcal{N}(0, 1) \tag{A.13}$$

$$\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2) \tag{A.14}$$

$$y_t = -0.05 y_{t-1} + 0.25 y_{t-2} - 0.63 y_{t-3} \tag{A.15}$$

$$- 4 \times 10^{-3} y_{t-1}^2 - 0.02 y_{t-3}^2 - 0.05 y_{t-2}^2 \tag{A.16}$$

$$- 0.048 y_{t-2} y_{t-1} - 0.02 y_{t-1} y_{t-3} - 0.06 y_{t-2} y_{t-3} \tag{A.17}$$

$$+ \varepsilon \tag{A.18}$$

The initial state y_0 is sampled from the standard Gaussian distribution. At time t, y_t is determined by a series expansion up to degree two, of monomials constructed from $y_{t-1}, y_{t-2}, y_{t-3}$, plus Gaussian distributed noise ε.

Using the sampling procedure described above, we construct 10 sample paths of temporal length 100 for training and testing the GP-AR model (some of them shown in figure A.1). The portion $t > 40$ of the black path in figure A.1 is reserved for model evaluation while its earlier segment as well as all other sample paths are retained for model training.

We construct a GP-AR model with input $\mathbf{x}_t = \begin{bmatrix} y_{t-1} \\ y_{t-2} \\ y_{t-3} \end{bmatrix}$, as

$$\begin{aligned} y_t &= f(\mathbf{x}_t) + \epsilon \\ f(\mathbf{x}_t) &\sim \mathcal{GP}(m(\mathbf{x}_t), K(\mathbf{x}_t, \mathbf{x}_s)) \\ \epsilon &\sim \mathcal{N}(0, \sigma_\epsilon^2) \end{aligned}$$

with mean and covariance functions are defined as

$$\begin{aligned} m(\mathbf{x}_t) &= -0.05 y_{t-1} + 0.25 y_{t-2} - 0.63 y_{t-3} \\ K(\mathbf{x}_t, \mathbf{x}_s) &= \sigma^2 \exp\left(-\frac{1}{2}\frac{||\mathbf{x}_t - \mathbf{x}_s||^2}{l^2}\right) \end{aligned}$$

where the kernel and noise parameters are initialised as follows $\sigma^2 = 2$, $l = 1$, $\sigma_\epsilon^2 = 0.5$. In figure A.2, we show some samples drawn from this GP-AR prior. During the training procedure, the model's hyper-parameters are selected using the CSA procedure described in section 4.2.3. After choosing the model hyper-parameters, the model is tested on the portion $t > 40$ of the black path in figure A.1.

In figure A.3, the posterior predictive distribution of the GP-AR model is used to generate one step ahead predictions with error bars. The GP-AR model is able to approximate the quadratic dynamics of equation (A.13), although the prior mean function $m(\mathbf{x}_t) = -0.05 y_{t-1} + 0.25 y_{t-2} - 0.63 y_{t-3}$ only captures the linear components of equation (A.11), the kernel enables modelling of the quadratic components.

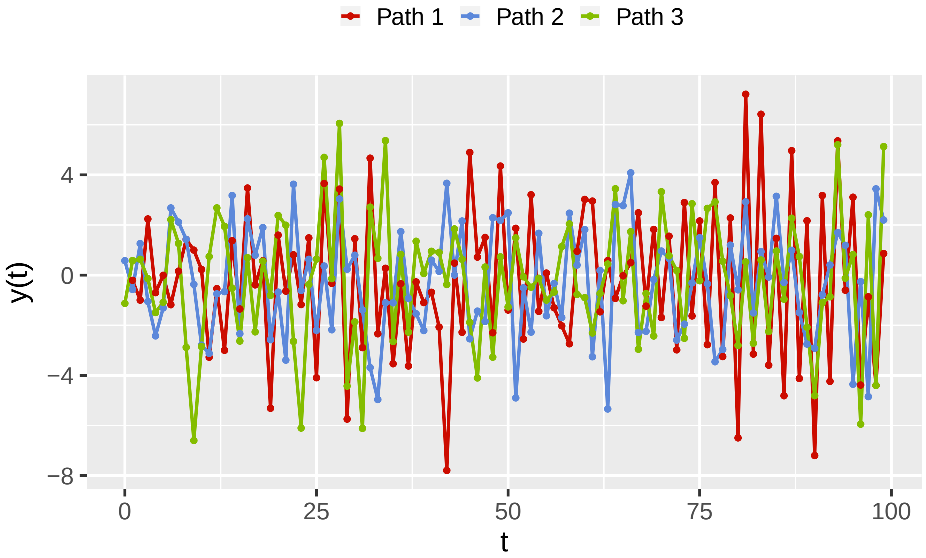


Figure A.2: Prior samples drawn from a GP-AR model (equations (A.8) and (A.9)) with an squared exponential kernel.

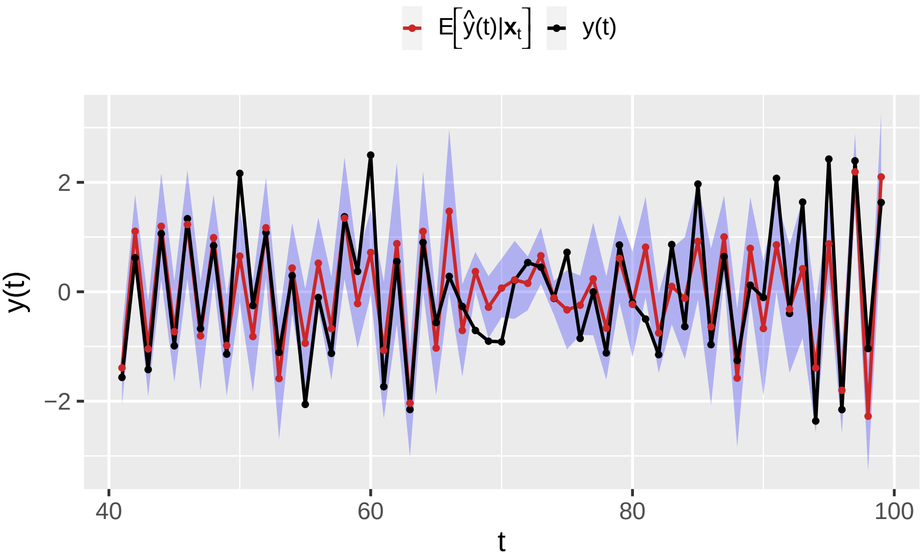


Figure A.3: Predictions made by the GP-AR model, error bars shaded in blue.

Appendix B

Sensitivity Analysis of the Radial Diffusion System

Model Sensitivity

Sensitivity analysis originated from the need to quantify the effect that small errors in parameters of a partial differential equation (PDE) have on its solution. Koda et al. [1979] described three approaches to sensitivity analysis of partial differential equations:

1. the direct method,
2. the variational method, and
3. the Fourier amplitude sensitivity test (FAST).

They compared these techniques by applying them on the problem of atmospheric diffusion.

Sensitivity of partial differential equations to their parameters has important implications for PDE constrained inverse problems. As a general rule, parameter identifiability increases with increasing sensitivity. For the Bayesian PDE inverse problem, this implies that posterior uncertainties of a parameter decrease if the forward model is sensitive to that parameter.

In this appendix, we apply the direct method of sensitivity analysis on the radial diffusion system from section 6.2 and visualise the sensitivity of the phase space density to the parameters of the diffusion field, the loss rate, and the source term.

Sensitivity: The Direct Method

Consider the PDE system introduced in equation (6.4)

$$\mathcal{L}_\theta f(x) = q_\theta(x) \ , \tag{B.1}$$

where $\mathcal{L}_\theta$ is a differential operator, $q_\theta(x)$ a source term, and θ is a vector containing all the parameters which define the forward model $\mathcal{F}_\theta = (\mathcal{L}_\theta, q_\theta)$. The sensitivity of the PDE in the region around $\theta = \theta^*$ is defined as the local gradient of its solution $f(\ell, t; \theta)$ with respect to the model parameters θ, which is given by

$$s(\ell, t, \theta) = \frac{\partial}{\partial \theta} f(\ell, t; \theta) \bigg|_{\theta=\theta^*} \ .$$

The sensitivity $s(\ell, t, \theta)$ is a function of space, time, and the forward model parameters. In the direct method of sensitivity analysis, it can be computed by differentiating both sides of equation (B.1) with respect to θ as follows:

$$\frac{\partial}{\partial \theta} \mathcal{L}_\theta[f(x)] = \frac{\partial}{\partial \theta} q_\theta(x) \ .$$

The expression above can be written as a PDE expressed in terms of the model sensitivity $s(\ell, t, \theta)$, the solution $f(\ell, t; \theta)$, and their derivatives, given by

$$\tilde{\mathcal{L}}_\theta[s(\ell, t, \theta), f(\ell, t; \theta)] = \frac{\partial}{\partial \theta} q_\theta(x) \ , \tag{B.2}$$

where $\tilde{\mathcal{L}}$ is the differential operator resulting from the differentiation of equation (B.1) by θ. Equations (B.1) and (B.2) form a system of coupled PDEs which can be solved numerically to obtain $f(\ell, t; \theta)$ and $s(\ell, t, \theta)$.

Sensitivity of the Phase Space Density

The direct method of sensitivity analysis can be applied on the radial diffusion model presented in section 6.2. Let $\theta = [\theta_\kappa, \theta_\lambda, \theta_q]$ be a vector containing all the parameters of the forward model, where θ_κ, θ_λ, and θ_q are the parameters of the diffusion field $\kappa(\ell, t)$, loss rate $\lambda(\ell, t)$, and source term $q(\ell, t)$ respectively. It is assumed that the parametrization scheme for $\kappa(\ell, t)$, $\lambda(\ell, t)$, and $q(\ell, t)$ is that which is outlined in section 6.2.1. Therefore, θ_κ, θ_λ, and θ_q are each three dimensional, making θ a nine dimensional vector.

Differentiating radial diffusion PDE in equation (6.1) with respect to an individual parameter θ_i yields the PDE describing the sensitivity component s_i, given by

$$\begin{aligned}\frac{\partial s_i}{\partial t} = \ell^2 \frac{\partial}{\partial \ell}&\left(\frac{\kappa(\ell,t)}{\ell^2}\frac{\partial s_i}{\partial \ell} + \frac{1}{\ell^2}\frac{\partial \kappa(\ell,t)}{\partial \theta_i}\frac{\partial f}{\partial \ell}\right)_{\mathcal{M},J} \\ &- \lambda(\ell,t) s_i - \frac{\partial \lambda(\ell,t)}{\partial \theta_i} f \\ &+ \frac{\partial q(\ell,t)}{\partial \theta_i}\ .\end{aligned} \tag{B.3}$$

We can simplify equation (B.3) further depending on which quantity the parameter θ_i belongs to. Equations (B.4) to (B.6) give the sensitivity equations for the parameters of $\kappa(\ell, t)$, $\lambda(\ell, t)$, and $q(\ell, t)$ respectively.

$$\frac{\partial s_\kappa}{\partial t} = \ell^2 \frac{\partial}{\partial \ell}\left(\frac{\kappa(\ell,t)}{\ell^2}\frac{\partial s_\kappa}{\partial \ell} + \frac{1}{\ell^2}\frac{\partial \kappa(\ell,t)}{\partial \theta_\kappa}\frac{\partial f}{\partial \ell}\right)_{\mathcal{M},J} - \lambda(\ell,t) s_\kappa \tag{B.4}$$

$$\frac{\partial s_\lambda}{\partial t} = \ell^2 \frac{\partial}{\partial \ell}\left(\frac{\kappa(\ell,t)}{\ell^2}\frac{\partial s_\lambda}{\partial \ell}\right)_{\mathcal{M},J} - \lambda(\ell,t) s_\lambda - \frac{\partial \lambda(\ell,t)}{\partial \theta_\lambda} f \tag{B.5}$$

$$\frac{\partial s_q}{\partial t} = \ell^2 \frac{\partial}{\partial \ell}\left(\frac{\kappa(\ell,t)}{\ell^2}\frac{\partial s_q}{\partial \ell}\right)_{\mathcal{M},J} - \lambda(\ell,t) s_q + \frac{\partial}{\partial \theta_q} q(\ell,t) \tag{B.6}$$

Example

We solve equations (6.1) and (B.4) to (B.6) numerically, using the boundary conditions stated in section 6.5.1. The PSD and the sensitivities are computed on a space-time grid of size 50×100, for the values listed in table 6.2.

Figure B.1 shows the computed solution for the PSD as a function of time on the x-axis and L^* or ℓ on the y-axis. The parameters sensitivities can differ by an order of magnitude or more, hence we plot them on separate charts for greater clarity.

From figures B.2 and B.3 we observe that the forward model is most sensitive to the parameter b. In fact within each parameter set θ_κ, θ_λ, and θ_q, the model is more sensitive to parameter b as compared to α and β.

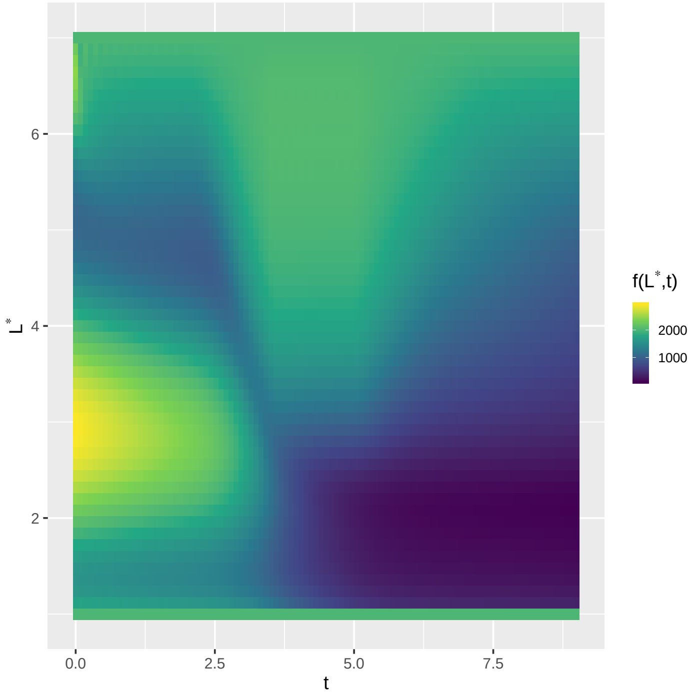


Figure B.1: The radial diffusion solution.

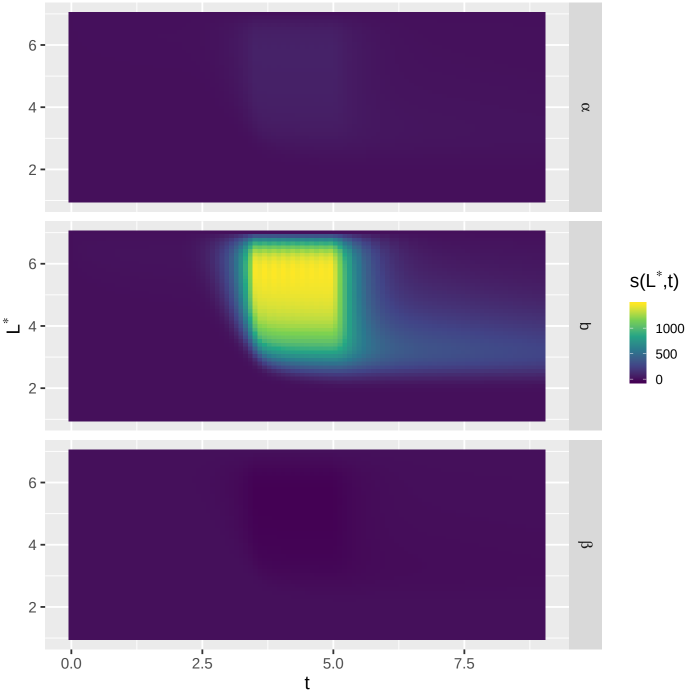


Figure B.2: Sensitivity of the forward model to the parameters of the source term.

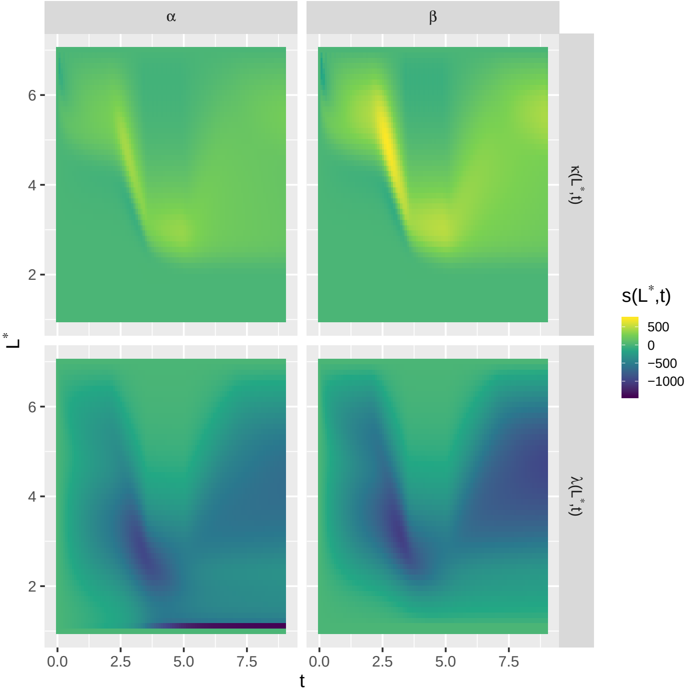


Figure B.3: Sensitivity of the forward model to the parameters α and β of the diffusion field and loss rate.

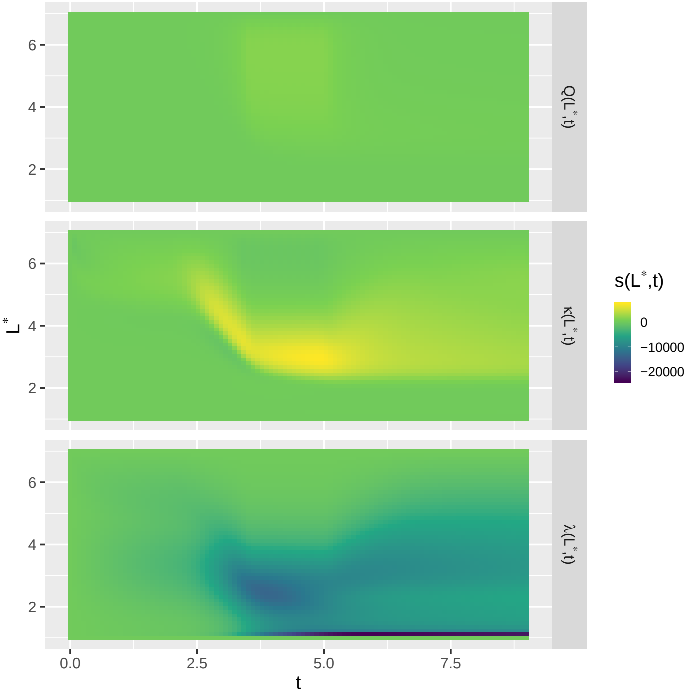


Figure B.4: Sensitivity of the forward model to the parameter b of the diffusion field, loss rate, and the source term.

Appendix C

Log Likelihood Of The DTLR Model

Direct computation

Due to the single effect constraint in equation (7.3) the mixture model in equation (7.4) can be expressed as

$$
\begin{aligned}
P(\mathbf{y}|\mathbf{x}) &= \Big(\sum_{i\in T} \hat{p}_i(\mathbf{x}) \prod_{j\in T} \sqrt{\frac{1+\alpha_{ji}}{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(1+\alpha_{ji})\big(y_j-\hat{y}_j(\mathbf{x})\big)^2}\Big) \\
&= \Big(\sum_{i\in T} \hat{p}_i(\mathbf{x}) \prod_{j\in T} \sqrt{\frac{1+\alpha_{ji}}{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}\alpha_{ji}\big(y_j-\hat{y}_j(\mathbf{x})\big)^2}\Big) \\
&\times \ \exp\Big(-\frac{1}{2\sigma^2}\sum_{j\in T}\big(y_j-\hat{y}_j(\mathbf{x})\big)^2\Big).
\end{aligned}
$$

Let $\theta \overset{\text{def}}{=} (\hat{\mathbf{y}}, \hat{\mathbf{p}}, \sigma, \alpha)$ denote the parameters of the model and consider the probability that $\hat{y}_i$ is the predictor corresponding to the ground truth time-lag, conditioned on the pair $(\mathbf{x}, \mathbf{y})$:

$$
\begin{aligned}
q_i(\mathbf{x},\mathbf{y}) &= P(\tau_i = 1|x,\mathbf{y}) \\
&= \frac{\hat{p}_i(\mathbf{x}) \exp\Big(-\frac{1}{2\sigma^2}\sum_{j\in T}\alpha_{ji}\big(y_j-\hat{y}_j(\mathbf{x})\big)^2 + \frac{1}{2}\sum_{j\in T}\log(1+\alpha_{ji})\Big)}{Z(\mathbf{x},\mathbf{y}|\theta)},
\end{aligned}
$$

with

$$Z(\mathbf{x},\mathbf{y}|\theta) = \sum_{i\in T} \hat{p}_i(\mathbf{x}) \exp\Big(-\frac{1}{2\sigma^2}\sum_{j\in T}\alpha_{ji}\big(y_j - \hat{y}_j(\mathbf{x})\big)^2 + \frac{1}{2}\sum_{j\in T}\log(1+\alpha_{ji})\Big).$$

This results in

$$\begin{aligned}\mathcal{L}[\{(\mathbf{x},\mathbf{y})\}_{\text{data}}|\theta] = &-|T|\log(\sigma)\\ &-\mathbb{E}_{\text{data}}\Big[\sum_{i\in T}\frac{1}{2\sigma^2}\big(y_i - \hat{y}_i(\mathbf{x})\big)^2 - \log\big(Z(\mathbf{x},\mathbf{y}|\theta)\big)\Big].\end{aligned}$$

Large deviation argument

Even though the log likelihood can be obtained by direct summation, for sake of generality we show how this can result from a large deviation principle. Assume that the number of learning samples tends to infinity, and so that in a small volume $dv = d\mathbf{x}d\mathbf{y}$ around a given joint configuration $(\mathbf{x},\mathbf{y})$, the number of data $N_{\mathbf{x},\mathbf{y}}$ becomes large. Restricting the likelihood to this subset of the data yields the following:

$$\mathcal{L}_{\mathbf{x},\mathbf{y}} = \prod_{m=1}^{N_{\mathbf{x},\mathbf{y}}} \sum_{\{\tau^{(m)}\}} \frac{\hat{p}(\tau^{(m)}|\mathbf{x})}{\prod_{i\in T}\sqrt{2\pi}\ \sigma_i(\tau^{(m)})} \exp\Big(-\frac{1}{2}\sum_{i\in T}\frac{\big(y_i - \hat{y}_i(\mathbf{x})\big)^2}{\sigma_i(\tau^{(m)})^2}\Big).$$

Upon introducing the relative frequencies:

$$q_i(\mathbf{x},\mathbf{y}) = \frac{1}{N_{\mathbf{x},\mathbf{y}}}\sum_{m=1}^{N_{\mathbf{x},\mathbf{y}}}\tau_i^{(m)} \qquad \text{satisfying} \qquad \sum_{i\in T} q_i(\mathbf{x},\mathbf{y}) = 1,$$

the sum over the $\tau_i^{(m)}$ is replaced by a sum over these new variables, with the summand obeying a large deviation principle

$$\mathcal{L}_{\mathbf{x},\mathbf{y}} \asymp \sum_{\mathbf{q}} \exp\Big(-N_{\mathbf{x},\mathbf{y}}\mathcal{F}_{x,\mathbf{y}}[\mathbf{q}]\Big)$$

where the rate function reads

$$\begin{aligned}\mathcal{F}_{\mathbf{x},\mathbf{y}}\big[\mathbf{q}\big] =&|T|\log(\sigma)\\ &+\sum_{i\in T}\Big[\big(y_i-\hat{y}_i(\mathbf{x})\big)^2\frac{1+\sum_{j\in T}\alpha_{ij}q_j}{2\sigma^2}\\ &\qquad-\frac{1}{2}q_i\sum_{j\in T}\log(1+\alpha_{ji})+q_i\log\frac{q_i}{\hat{p}_i}\Big].\end{aligned}$$

Taking the saddle point for q_i yields equation (7.6). Inserting this into $\mathcal{F}$ and taking the average over the data set yields the log likelihood (equation (7.7)) with opposite sign:

$$\mathcal{L}[\{(\mathbf{x},\mathbf{y})\}_{\text{data}}|\theta] = -\mathbb{E}_{\text{data}}\Big[\mathcal{F}_{x,\mathbf{y}}\big[\mathbf{q}(\mathbf{x},\mathbf{y})\big]\Big].$$

Saddle point equations

Now we turn to the self-consistent equations relating the parameters θ of the model at a saddle point of the log likelihood function. First, the optimization of the predictors $\hat{\mathbf{y}}$ yields:

$$\frac{\partial\mathcal{L}}{\partial\hat{y}_i(\mathbf{x})} = \frac{1}{\sigma^2}\mathbb{E}_{\text{data}}\Big[\big(y_i-\hat{y}_i(\mathbf{x})\big)\big(1+\sum_{j\in T}\alpha_{ij}q_j(\mathbf{x},\mathbf{y})\big)\Big|\mathbf{x}\Big].$$

Then the optimization of $\hat{\mathbf{p}}$ gives:

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial\hat{p}_i(\mathbf{x})} &= \mathbb{E}_{\text{data}}\Big[\frac{q_i(\mathbf{x},\mathbf{y})}{\hat{p}_i(\mathbf{x})}-\lambda(\mathbf{x})\Big|\mathbf{x}\Big],\\ &= \frac{1}{\hat{p}_i(\mathbf{x})}\mathbb{E}_{\text{data}}\Big[q_i(\mathbf{x},\mathbf{y})\Big|\mathbf{x}\Big]-\lambda(\mathbf{x})\end{aligned}$$

with $\lambda(\mathbf{x})$ a Lagrange multiplier to insure that $\sum_i\hat{p}_i(\mathbf{x})=1$ This gives

$$\hat{p}_i(\mathbf{x}) = \frac{1}{\lambda(\mathbf{x})}\mathbb{E}_{\text{data}}\Big[q_i(\mathbf{x},\mathbf{y})\Big|\mathbf{x}\Big]$$

Hence

$$\sum_{i\in T}\hat{p}_i(\mathbf{x}) = \frac{1}{\lambda(\mathbf{x})} = 1 \qquad \forall x$$

in order to fulfill the normalization constraint, yielding

$$\hat{y}_i(x) = \frac{\mathbb{E}_{\text{data}}\Big[y_i\big(1 + \sum_{j \in T} \alpha_{ij} q_j(\mathbf{x}, \mathbf{y})\big)\Big|\mathbf{x}\Big]}{\mathbb{E}_{\text{data}}\Big[1 + \sum_{j \in T} \alpha_{ij} q_j(\mathbf{x}, \mathbf{y})\Big|\mathbf{x}\Big]} \tag{C.1}$$

$$\hat{p}_i(\mathbf{x}) = \mathbb{E}_{\text{data}}\Big[q_i(\mathbf{x}, \mathbf{y})\Big|\mathbf{x}\Big], \tag{C.2}$$

Finally the optimization of α reads:

$$\frac{\partial \mathcal{L}}{\partial \alpha_{ij}} = \frac{1}{2(1 + \alpha_{ij})} \mathbb{E}_{\text{data}}\big[q_j(\mathbf{x}, \mathbf{y})\big] - \frac{1}{2\sigma^2} \mathbb{E}_{\text{data}}\big[\big(y_i - \hat{y}_i(\mathbf{x})\big)^2 q_j(\mathbf{x}, \mathbf{y})\big].$$

Appendix D

Stability Analysis Of The DTLR Model

The analysis is restricted for simplicity to the case $\alpha_{ij} = \alpha\delta_{ij}$. The log likelihood as a function of $r = \sigma^2/\sigma_0^2$ and $\beta = \alpha/r$ after inserting the optimal $\mathbf{q} = \mathbf{q}(\mathbf{x}, \mathbf{y})$ reads in that case

$$\mathcal{L}(r, \beta) = -\frac{|T|}{2}\log(r) - \frac{|T|}{2r} + \frac{1}{2}\log(1 + r\beta) + \mathbb{E}_{\text{data}}\Big[\log(Z) - \lambda(\mathbf{x})\sum_{i\in T}\hat{p}_i(\mathbf{x})\Big]$$

with

$$Z = \sum_i \hat{p}_i(\mathbf{x})\exp\Big(-\frac{\beta}{2\sigma_0^2}\Delta y_i^2(\mathbf{x})\Big),$$

and where $\lambda(\mathbf{x})$ is a Lagrange multiplier which has been added to impose the normalization of $\hat{\mathbf{p}}$. The gradient reads

$$\begin{aligned}
\frac{\partial\mathcal{L}}{\partial r} &= \frac{1}{2r^2}\Big(|T|(1-r)+\frac{\beta r^2}{1+\beta r}\Big),\\
\frac{\partial\mathcal{L}}{\partial \beta} &= \frac{r}{2(1+r\beta)}-\frac{1}{2}C_1[\mathbf{q}],\\
\frac{\partial\mathcal{L}}{\partial \hat{y}_i(\mathbf{x})} &= \frac{1}{\sigma^2}\mathbb{E}_{\text{data}}\Big[\big(y_i-\hat{y}_i(\mathbf{x})\big)\big(1+\alpha q_i(\mathbf{x},\mathbf{y})\big)\Big|\mathbf{x}\Big].\\
\frac{\partial\mathcal{L}}{\partial \hat{p}_i(\mathbf{x})} &= \frac{\mathbb{E}_{\text{data}}\big[q_i(\mathbf{x},\mathbf{y})|\mathbf{x}\big]}{\hat{p}_i(\mathbf{x})}-\lambda(\mathbf{x}),
\end{aligned}$$

with

$$C_1[\mathbf{q}] = \frac{1}{\sigma_0^2}\mathbb{E}_{\text{data}}\Big(\sum_{i\in T} q_i(\mathbf{x},\mathbf{y})\Delta y_i^2(\mathbf{x})\Big),$$

This leads to the following relation at the saddle point:

$$\begin{aligned}
r &= \frac{|T|-C_1[\mathbf{q}]}{|T|-1},\\
\alpha &= \frac{|T|}{|T|-1}\frac{1-C_1[\mathbf{q}]}{C_1[\mathbf{q}]},\\
\hat{y}_i(\mathbf{x}) &= \frac{\mathbb{E}_{\text{data}}\Big[y_i\big(1+\alpha q_i(\mathbf{x},\mathbf{y})\big)\Big|\mathbf{x}\Big]}{\mathbb{E}_{\text{data}}\Big[1+\alpha q_i(\mathbf{x},\mathbf{y})\Big|\mathbf{x}\Big]}\\
\hat{p}_i(\mathbf{x}) &= \mathbb{E}_{\text{data}}\big[q_i(\mathbf{x},\mathbf{y})|\mathbf{x}\big].
\end{aligned}$$

Let us now compute the Hessian. It is easy to see that the block corresponding to the predictors $\hat{\mathbf{y}}$ decouples from the rest as soon as these predictors are centered.

Denoting

$$C_2[\mathbf{q}] = \frac{1}{\sigma_0^4}\mathbb{E}_{\text{data}}\Big[\sum_{i\in T} q_i(\mathbf{x},\mathbf{y})\Big(\Delta y_i^2(\mathbf{x})-\sum_{j=1}^{n} q_j(\mathbf{x},\mathbf{y})\Delta y_j^2(\mathbf{x})\Big)^2\Big],$$

we have

$$\begin{aligned}
\frac{\partial^2 \mathcal{L}}{\partial r^2} &= \frac{1}{2r^2}\Big(-|T| + 2\frac{|T|}{|T|-1}\big(C_1[\mathbf{q}]-1\big) - \beta^2 C_1^2[\mathbf{q}]\Big)\\
\frac{\partial^2 \mathcal{L}}{\partial r \partial \beta} &= \frac{1}{2r^2} C_1^2[\mathbf{q}]\\
\frac{\partial^2 \mathcal{L}}{\partial \beta^2} &= \frac{1}{4}\Big(C_2[\mathbf{q}] - 2C_1^2[\mathbf{q}]\Big)\\
\frac{\partial^2 \mathcal{L}}{\partial \hat{p}_i(\mathbf{x}) \partial \hat{p}_j(\mathbf{x})} &= -\frac{\mathbb{E}_{\text{data}}\big[q_i(\mathbf{x},\mathbf{y}) q_j(\mathbf{x},\mathbf{y}) | \mathbf{x}\big]}{\hat{p}_i(\mathbf{x})\hat{p}_j(\mathbf{x})}\\
\frac{\partial^2 \mathcal{L}}{\partial r \partial \hat{p}_i(\mathbf{x})} &= 0\\
\frac{\partial^2 \mathcal{L}}{\partial \beta \partial \hat{p}_i(\mathbf{x})} &= -\frac{u_i[x,\mathbf{q}]}{2\hat{p}_i(\mathbf{x})},
\end{aligned}$$

where

$$u_i[x,\mathbf{q}] \overset{\text{def}}{=} \frac{1}{\sigma_0^2}\mathbb{E}_{\text{data}}\Big[q_i(\mathbf{x},\mathbf{y})\big(\Delta y_i^2(\mathbf{x}) - \sum_{j\in T} q_j(\mathbf{x},\mathbf{y})\Delta y_j^2(\mathbf{x})\big)|\mathbf{x}\Big].$$

There are two blocks in this Hessian, the one corresponding to r and β and the one corresponding to derivatives with respect to $\hat{p}_i$. The stability of the first one depends on the sign of $C_2[\mathbf{q}] - 2C_1^2[\mathbf{q}]$ for $|T|$ large while the second block is always stable as being an average of the exterior product of the vector $(q_1(\mathbf{x},\mathbf{y})/\hat{p}_1(\mathbf{x}),\ldots,q_{|T|}(\mathbf{x},\mathbf{y})/\hat{p}_{|T|}(\mathbf{x}))$ by itself. At the degenerate point $\alpha = 0$, $r = 1$, $\hat{p}_i = 1/|T|$ the Hessian simplifies as follows. Denote

$$d\eta = dr\mathbf{e}_1 + d\beta\mathbf{e}_2 + \int d\mathbf{x} \sum_{i\in T} d\hat{p}_i(\mathbf{x})\mathbf{e}_{i+2}(\mathbf{x})$$

a given vector of perturbations, decomposed onto a set of unit tangent vectors, $\{\mathbf{e}_1$ and $\mathbf{e}_2\}$ being respectively associated to r and β, while $\mathbf{e}_i(\mathbf{x})$

associated to $\hat{p}_i(\mathbf{x})$ for all $i \in T$ and $x \in \mathcal{X}$. Denote

$$\mathbf{u} = \sum_{i \in T} \int d\mathbf{x} u_i[\mathbf{x}] \mathbf{e}_i(\mathbf{x})$$

$$\mathbf{v}(\mathbf{x}) = \sum_{i \in T} \mathbf{e}_i(\mathbf{x})$$

with

$$C_2 = \frac{1}{|T|\sigma_0^4} \mathbb{E}_{\text{data}} \Big[\sum_{i \in T} \Big(\Delta y_i^2(\mathbf{x}) - \frac{1}{|T|} \sum_{j \in T} \Delta y_j^2(\mathbf{x}) \Big)^2 \Big].$$

$$u_i[\mathbf{x}] = \frac{1}{\sigma_0^2} \mathbb{E}_{\text{data}} \big[\Delta y_i^2(\mathbf{x}) - \sigma_0^2 | \mathbf{x} \big].$$

With these notations the Hessian reads:

$$H = \frac{1}{2} \Big(-|T| \mathbf{e}_1 \mathbf{e}_1^t + \mathbf{e}_1 \mathbf{e}_2^t + \mathbf{e}_2 \mathbf{e}_1^t + \big(\frac{C_2}{2} - 1 \big) \mathbf{e}_2 \mathbf{e}_2^t - \mathbf{u} \mathbf{e}_2^t - \mathbf{e}_2 \mathbf{u}^t - \int d\mathbf{x} \mathbf{v}(\mathbf{x}) \mathbf{v}^t(\mathbf{x}) \Big).$$

In fact we are interested in the eigenvalues of H in the subspace of deformations which conserve the norm of $\hat{\mathbf{p}}$, i.e. orthogonal to $\mathbf{v}(\mathbf{x})$, thereby given by

$$\eta = \eta_1 \mathbf{e}_1 + \eta_2 \mathbf{e}_2 + \eta_3 \mathbf{u}.$$

In this subspace the Hessian reads

$$H = \frac{1}{2} \begin{bmatrix} -|T| & 1 & 0 \\ 1 & \frac{C_2}{2} - 1 & -M|T|C_2 \\ 0 & -M|T|C_2 & 0 \end{bmatrix},$$

where M is the number of data points, resulting from the fact that

$$\sum_{i \in T} \int d\mathbf{x} u_i[\mathbf{x}]^2 = \frac{M}{\sigma_0^4} \mathbb{E}_{\text{data}} \Big[\sum_{i \in T} \big(\Delta y_i^2(\mathbf{x}) - \sigma_0^2 \big)^2 \Big],$$

$$= MC_2,$$

because $\mathbb{E}_{\text{data}}(\cdot|\mathbf{x})$ as a function of x is actually a point-wise function on the data. If $|u|^2 > 0$ or if $|u| = 0$ and $1 + |T|(C_2/2 - 1) > 0$ there is at least one positive eigenvalue. Let Λ be such an eigenvalue. After eliminating dr and $d\beta$ from the eigenvalue equations in $d\eta$, the deformation along this mode verifies

$$d\eta \propto \Lambda \mathbf{e}_1 + \Lambda(|T| + \Lambda)\mathbf{e}_2 - M|T|(|T| + \Lambda)C_2\mathbf{u},$$

which corresponds to increasing r and α while decreasing for each x the $\hat{p}_i$ having the highest mean relative error $u_i[\mathbf{x}]$.
Concerning solutions for which

$$\hat{p}_i(\mathbf{x}) = \delta_{i\hat{I}(\mathbf{x})}$$

is concentrated on some index $\hat{I}(\mathbf{x})$, the analysis is more complex. In that case $C_2[\mathbf{p}] = 0$ and $C_1[\mathbf{p}] > 0$. The (r, β) sector has 2 negative eigenvalues, while the $\hat{\mathbf{p}}$ block is $(-)$ a covariance matrix, so it has as well negative eigenvalues. The coupling between these two blocks could however in principle generate in some cases some instabilities.

Still, the log likelihood of such solutions reads

$$\mathcal{L} = -\frac{|T|}{2}\log(\sigma^2) + \frac{1}{2}\log(1+\alpha) - \frac{1}{2\sigma^2}\mathbb{E}_{\text{data}}\Big[\sum_{i\in T}\Delta y_i^2(\mathbf{x})\Big] - \frac{\alpha}{2\sigma^2}\mathbb{E}_{\text{data}}\Big[\Delta y_{I(\mathbf{x})}^2(\mathbf{x})\Big]$$

so we get the following optimal solution

$$\sigma^2 = \frac{1}{|T|}\mathbb{E}_{\text{data}}\Big[\sum_{i\in T}\Delta y_i^2(\mathbf{x})\Big],$$

$$\frac{1}{1+\alpha} = \frac{\mathbb{E}_{\text{data}}\Big[\Delta y_{I(\mathbf{x})}^2(\mathbf{x})\Big]}{\sigma^2},$$

$$I(\mathbf{x}) = \underset{i\in T}{\operatorname{argmin}}\,\mathbb{E}_{\text{data}}\Big[\Delta y_i^2(\mathbf{x})|\mathbf{x}\Big].$$

Appendix E

The Optimal DTLR Predictor

Given $I(\mathbf{x})$ a candidate index function we associate the point-like measure

$$p_i(\mathbf{x}) = \delta_{i,I(\mathbf{x})}.$$

Written in terms of p the loss function reads

$$\mathcal{L}_2(\hat{y}, p) = \mathbb{E}_{x,\mathbf{y}}\Big[\sum_{i \in T} p_i(\mathbf{x})\big(y_i - \hat{y}(\mathbf{x})\big)^2\Big].$$

Under equation (7.4) (with $\alpha_{ij} = \alpha\delta_{ij}$) the loss is equal to

$$\mathcal{L}_2(\hat{y}, p) = \mathbb{E}_x\Big[\sum_{i \in T} p_i(\mathbf{x})\Big(\big(\hat{y}_i(\mathbf{x}) - \hat{y}(\mathbf{x})\big)^2 - \hat{p}_i(\mathbf{x})\frac{\alpha\sigma^2}{1+\alpha}\Big)\Big] + \sigma^2$$

The minimization w.r.t. $\hat{y}$ yields

$$\hat{y}(\mathbf{x}) = \sum_{i \in T} p_i(\mathbf{x})\hat{y}_i(\mathbf{x}). \tag{E.1}$$

In turn, as a function of p_i the loss being a convex combination, its minimization yields

$$p_i(\mathbf{x}) = \delta_{i,I(\mathbf{x})}, \tag{E.2}$$

$$I(\mathbf{x}) = \operatorname*{argmin}_{i \in T}\Big(\big(\hat{y}_i(\mathbf{x}) - \hat{y}(\mathbf{x})\big)^2 - \hat{p}_i(\mathbf{x})\frac{\alpha\sigma^2}{1+\alpha}\Big). \tag{E.3}$$

Combining these equations (E.1,E.2,E.3) we get

$$I(\mathbf{x}) = \underset{i \in T}{\operatorname{argmax}}\big(\hat{p}_i(\mathbf{x})\big),$$

which concludes the proof.

Bibliography

Lucie P. Aarts and Peter van der Veer. Neural network method for solving partial differential equations. *Neural Processing Letters*, 14(3):261–271, Dec 2001. ISSN 1573-773X. doi: 10.1023/A:1012784129883. URL `https://doi.org/10.1023/A:1012784129883`. 93

M Abramowitz and I Stegun. *A (1972) Handbook of Mathematical Functions (3rd Edition) Dover*. 1972. 98

S.-I. Akasofu. Energy coupling between the solar wind and the magnetosphere. *Space Science Reviews*, 28:121–190, 1981. doi: 10.1007/BF00218810. 61

H. Alfvén. Magneto hydrodynamic waves, and the heating of the solar corona. *Monthly Notices of the Royal Astronomical Society*, 107:211, 1947. doi: 10.1093/mnras/107.2.211. 18

Hannes Alfvén. Existence of electromagnetic-hydrodynamic waves. *Nature*, 150(3805):405, 1942. 19

Hannes Alfvén. On the existence of electromagnetic-hydrodynamic waves. *Arkiv for matematik, astronomi och fysik*, 29:1–7, 1943. 19

Martin D Altschuler and Gordon Newkirk. Magnetic fields and the structure of the solar corona. *Solar Physics*, 9(1):131–149, 1969. 18, 113

A. Aminataei and M.M. Mazarei. Numerical solution of poisson's equation using radial basis function networks on the polar coordinate. *Computers & Mathematics with Applications*, 56(11):2887 – 2895, 2008. ISSN 0898-1221. doi: https://doi.org/10.1016/j.camwa.2008.07.026. URL `http://www.sciencedirect.com/science/article/pii/S0898122108004616`. 93

CN Arge and VJ Pizzo. Improvement in the prediction of solar wind conditions using near-real time solar magnetic field updates. *Journal of Geophysical Research: Space Physics*, 105(A5):10465–10479, 2000. 113, 114

CN Arge, JG Luhmann, D Odstrcil, CJ Schrijver, and Y Li. Stream structure and coronal sources of the solar wind during the may 12th, 1997 cme. *Journal of Atmospheric and Solar-Terrestrial Physics*, 66(15-16): 1295–1309, 2004. 113

E Astafyeva, Yu Yasyukevich, A Maksikov, and I Zhivetiev. Geomagnetic storms, super-storms, and their impacts on gps-based navigation systems. *Space Weather*, 12(7):508–525, 2014. 63

Daniel N Baker. How to cope with space weather. *Science*, 297(5586): 1486–1487, 2002. 88

DN Baker, Eamonn Daly, Ioannis Daglis, John G Kappenman, and Mikhail Panasyuk. Effects of space weather on technology infrastructure. *Space Weather*, 2(2), 2004. 10

Ramkumar Bala and Patricia Reiff. Improvements in short-term forecasting of geomagnetic activity. *Space Weather*, 10(6):1–11, 2012. ISSN 15427390. doi: 10.1029/2012SW000779. xv, 63, 65, 75

Ramkumar Bala, P. H. Reiff, and J. E. Landivar. Real-time prediction of magnetospheric activity using the boyle index. *Space Weather*, 7(4): n/a–n/a, 2009. ISSN 1542-7390. doi: 10.1029/2008SW000407. URL `http://dx.doi.org/10.1029/2008SW000407`. 35

M. A. Balikhin, O. M. Boaghe, S. A. Billings, and H. St C. K. Alleyne. Terrestrial magnetosphere as a nonlinear resonator. *Geophysical Research Letters*, 28(6):1123–1126, 2001. ISSN 1944-8007. doi: 10.1029/2000GL000112. URL `http://dx.doi.org/10.1029/2000GL000112`. 34

P. Ballatore and W. D. Gonzalez. On the estimates of the ring current injection and decay. *Earth, Planets and Space*, 55(7):427–435, 2014. ISSN 1880-5981. doi: 10.1186/BF03351776. URL `http://dx.doi.org/10.1186/BF03351776`. 34

J. Bartels and J. Veldkamp. International data on magnetic disturbances, second quarter, 1949. *Journal of Geophysical Research*, 54(4):399–400, 1949. ISSN 2156-2202. doi: 10.1029/JZ054i004p00399. URL `http://dx.doi.org/10.1029/JZ054i004p00399`. 33

J. Bartels, N. H. Heck, and H. F. Johnston. The three-hour-range index measuring geomagnetic activity. *Terrestrial Magnetism and Atmospheric Electricity (Journal of Geophysical Research)*, 44:411, 1939. doi: 10.1029/TE044i004p00411. 90

S. Basu, W. J. Chaplin, Y. Elsworth, R. New, and A. M. Serenelli. Fresh Insights on the Structure of the Solar Core. *The Astrophysical Journal*, 699:1403–1417, July 2009. doi: 10.1088/0004-637X/699/2/1403. 16

M. Baymani, S. Effati, and A. Kerayechian. A feed-forward neural network for solving stokes problem. *Acta Applicandae Mathematicae*, 116(1):55, Jul 2011. ISSN 1572-9036. doi: 10.1007/s10440-011-9627-5. URL `https://doi.org/10.1007/s10440-011-9627-5`. 93

Yoshua Bengio, Patrice Simard, Paolo Frasconi, et al. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2):157–166, 1994. 66

Alain Berlinet and Christine Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer US, 2004. ISBN 978-1-4613-4792-7. doi: 10.1007/978-1-4419-9096-9. URL `http://www.springer.com/us/book/9781402076794`. 41, 141

L. Biermann. Kometenschweife und solare Korpuskularstrahlung. *Zeitschrift fur Astrophysik*, 29:274, Jan 1951. 18

L. Biermann. Über den Schweif des Kometen Halley im Jahre 1910. *Zeitschrift Naturforschung Teil A*, 7(1):127–136, Jan 1952. doi: 10.1515/zna-1952-0122. 18

L. Biermann. Solar corpuscular radiation and the interplanetary gas. *The Observatory*, 77:109–110, Jun 1957. 18

S. A. Billings, S. Chen, and M. J. Korenberg. Identification of mimo non-linear systems using a forward-regression orthogonal estimator.

International Journal of Control, 49(6):2157–2189, 1989. doi: 10.1080/00207178908559767. URL `http://www.tandfonline.com/doi/abs/10.1080/00207178908559767`. 34

Stephen A Billings. *Nonlinear system identification: NARMAX methods in the time, frequency, and spatio-temporal domains*. John Wiley & Sons, 2013. 34

Space Studies Board. Committee on the societal and economic impacts of severe space weather events: A workshop. 2008. 8

Space Studies Board, National Research Council, et al. *Severe space weather events: Understanding societal and economic impacts: A workshop report*. National Academies Press, 2009. 10

Volker Bothmer and Ioannis A Daglis. *Space weather: physics and effects*. Springer Science & Business Media, 2007. 10

CB Boyle, PH Reiff, and MR Hairston. Empirical polar cap potentials. *Journal of Geophysical Research*, 102(A1):111–125, 1997. 35

R. J. Boynton, M. A. Balikhin, S. A. Billings, A. S. Sharma, and O. A. Amariutei. Data derived narmax dst model. *Annales Geophysicae*, 29 (6):965–971, 2011a. doi: 10.5194/angeo-29-965-2011. URL `http://www.ann-geophys.net/29/965/2011/`. 34

R. J. Boynton, M. A. Balikhin, S. A. Billings, H. L. Wei, and N. Ganushkina. Using the narmax ols-err algorithm to obtain the most influential coupling functions that affect the evolution of the magnetosphere. *Journal of Geophysical Research: Space Physics*, 116(A5):n/a–n/a, 2011b. ISSN 2156-2202. doi: 10.1029/2010JA015505. URL `http://dx.doi.org/10.1029/2010JA015505`. 34

R. J. Boynton, M. A. Balikhin, S. A. Billings, G. D. Reeves, N. Ganushkina, M. Gedalin, O. A. Amariutei, J. E. Borovsky, and S. N. Walker. The analysis of electron fluxes at geosynchronous orbit employing a narmax approach. *Journal of Geophysical Research: Space Physics*, 118(4):1500–1513, 2013. ISSN 2169-9402. doi: 10.1002/jgra.50192. URL `http://dx.doi.org/10.1002/jgra.50192`. 34

D. H. Brautigam and J. M. Albert. Radial diffusion analysis of outer radiation belt electrons during the october 9, 1990, magnetic storm. *Journal of Geophysical Research: Space Physics*, 105(A1):291–309, 2000. ISSN 2156-2202. doi: 10.1029/1999JA900344. URL `http://dx.doi.org/10.1029/1999JA900344`. 90, 99, 100

Encyclopaedia Britannica. Great red spot. URL `https://www.britannica.com/place/Great-Red-Spot`. 6

R. K. Burton, R. L. McPherron, and C. T. Russell. An empirical relationship between interplanetary conditions and dst. *Journal of Geophysical Research*, 80(31):4204–4214, 1975. ISSN 2156-2202. doi: 10.1029/JA080i031p04204. URL `http://dx.doi.org/10.1029/JA080i031p04204`. 34, 61, 62

E. Camporeale, S. Wing, J. Johnson, C. M. Jackman, and R. McGranaghan. Space weather in the machine learning era: A multidisciplinary approach. *Space Weather*, 16(1):2–4, 2018a. doi: 10.1002/2017SW001775. URL `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2017SW001775`. 112

Enrico Camporeale. The Challenge of Machine Learning in Space Weather Nowcasting and Forecasting. *arXiv e-prints*, art. arXiv:1903.05192, Mar 2019. 112

Enrico Camporeale, Yuri Shprits, Mandar Chandorkar, Alexander Drozdov, and Simon Wing. On the propagation of uncertainties in radiation belt simulations. *Space Weather*, 14(11):982–992, 2016.

Enrico Camporeale, Simon Wing, and Jay Johnson. *Machine learning techniques for space weather*. Elsevier, 2018b. 112

Paul Cannon, Matthew Angling, Leslie Barclay, Charles Curry, Clive Dyer, Robert Edwards, Graham Greene, Mike Hapgood, Richard B Horne, David Jackson, et al. *Extreme space weather: impacts on engineered systems and infrastructure*. Royal Academy of Engineering, 2013. 10

Michael J Carlowicz and Ramon E Lopez. *Storms from the sun: the emerging science of space weather*. Joseph Henry Press, 2002. 8

M. Chandorkar, E. Camporeale, and S. Wing. Probabilistic forecasting of the disturbance storm time index: An autoregressive gaussian process approach. *Space Weather*, 15(8):1004–1019, Aug 2017. ISSN 1542-7390. doi: 10.1002/2017SW001627. URL `https://agupubs.onlinelibrary.wiley.com/doi/full/10.1002/2017SW001627`. 63, 64

Mandar Chandorkar and Enrico Camporeale. Chapter 9 - probabilistic forecasting of geomagnetic indices using gaussian process models. In Enrico Camporeale, Simon Wing, and Jay R. Johnson, editors, *Machine Learning Techniques for Space Weather*, pages 237 – 258. Elsevier, 2018. ISBN 978-0-12-811788-0. doi: 10.1016/B978-0-12-811788-0.00009-3. URL `http://www.sciencedirect.com/science/article/pii/B9780128117880000093`.

Stuart Clark and Stuart G Clark. *The Sun Kings: the unexpected tragedy of Richard Carrington and the tale of how modern astronomy began*. Princeton University Press, 2007. 8

Edward W Cliver and William F Dietrich. The 1859 space weather event revisited: limits of extreme activity. *Journal of Space Weather and Space Climate*, 3:A31, 2013. 18

Jon Cockayne, Chris Oates, Tim Sullivan, and Mark Girolami. Probabilistic numerical methods for pde-constrained bayesian inverse problems. *AIP Conference Proceedings*, 1853(1):060001, 2017. doi: 10.1063/1.4985359. URL `https://aip.scitation.org/doi/abs/10.1063/1.4985359`. 94

Patrick R Conrad, Mark Girolami, Simo Särkkä, Andrew Stuart, and Konstantinos Zygalakis. Statistical analysis of differential equations: introducing probability measures on numerical solutions. *Statistics and Computing*, 27(4):1065–1082, 2017. 94

Freddy Moreno Cárdenas, Sergio Cristancho Sánchez, and Santiago Vargas Domínguez. The grand aurorae borealis seen in colombia in 1859. *Advances in Space Research*, 57(1):257 – 267, 2016. ISSN 0273-1177. doi: https://doi.org/10.1016/j.asr.2015.08.026. URL `http://www.sciencedirect.com/science/article/pii/S0273117715006109`. 8

T. Neil Davis and Masahisa Sugiura. Auroral electrojet activity index ae and its universal time variations. *Journal of Geophysical Research*, 71(3): 785–801, 1966. ISSN 2156-2202. doi: 10.1029/JZ071i003p00785. URL `http://dx.doi.org/10.1029/JZ071i003p00785`. 33

J. De Keyser, M. W. Dunlop, C. J. Owen, B. U. Ö Sonnerup, S. E. Haaland, A. Vaivads, G. Paschmann, R. Lundin, and L. Rezeau. Magnetopause and boundary layer. *Space Science Reviews*, 118(1):231–320, Jun 2005. ISSN 1572-9672. doi: 10.1007/s11214-005-3834-1. URL `https://doi.org/10.1007/s11214-005-3834-1`. xiv, 24

A. J. Dessler and E. N. Parker. Hydromagnetic theory of geomagnetic storms. *Journal of Geophysical Research*, 64(12):2239–2252, 1959. ISSN 2156-2202. doi: 10.1029/JZ064i012p02239. URL `http://dx.doi.org/10.1029/JZ064i012p02239`. 33

Sander Dieleman, Jan Schlüter, Colin Raffel, Eben Olson, Søren Kaae Sønderby, Daniel Nouri, et al. Lasagne: First release., August 2015. URL `http://dx.doi.org/10.5281/zenodo.27878`. 69

Frank Dondelinger, Dirk Husmeier, Simon Rogers, and Maurizio Filippone. Ode parameter inference using adaptive gradient matching with gaussian processes. In Carlos M. Carvalho and Pradeep Ravikumar, editors, *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, volume 31 of *Proceedings of Machine Learning Research*, pages 216–228, Scottsdale, Arizona, USA, 29 Apr–01 May 2013. PMLR. URL `http://proceedings.mlr.press/v31/dondelinger13a.html`. 92

Yong Duan. A note on the meshless method using radial basis functions. *Computers & Mathematics with Applications*, 55(1):66 – 75, 2008. ISSN 0898-1221. doi: https://doi.org/10.1016/j.camwa.2007.03.011. URL `http://www.sciencedirect.com/science/article/pii/S089812210700332X`. 93

Yong Duan and Yong-Ji Tan. A meshless galerkin method for dirichlet problems using radial basis functions. *Journal of Computational and Applied Mathematics*, 196(2):394 – 401, 2006. ISSN 0377-0427. doi: https:

//doi.org/10.1016/j.cam.2005.09.018. URL `http://www.sciencedirect.com/science/article/pii/S0377042705005832`. 93

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159, 2011. 69

Ian T Durham. Rethinking the history of solar wind studies: Eddington's analysis of comet morehouse. *Notes and Records of the Royal Society*, 60(3):261–270, 2006. doi: 10.1098/rsnr.2006.0149. URL `https://royalsocietypublishing.org/doi/abs/10.1098/rsnr.2006.0149`. 18

Russell Eberhart and James Kennedy. A new optimizer using particle swarm theory. In *MHS'95. Proceedings of the Sixth International Symposium on Micro Machine and Human Science*, pages 39–43. Ieee, 1995. 63

M. Elansari, D. Ouazar, and A. H.-D. Cheng. Boundary solution of poisson's equation using radial basis function collocated on gaussian quadrature nodes. *Communications in Numerical Methods in Engineering*, 17(7):455–464, 2001. ISSN 1099-0887. doi: 10.1002/cnm.419. URL `http://dx.doi.org/10.1002/cnm.419`. 93

Jeffrey L. Elman. Finding structure in time. *Cognitive Science*, 14 (2):179–211, 1990. doi: 10.1207/s15516709cog1402_1. URL `https://onlinelibrary.wiley.com/doi/abs/10.1207/s15516709cog1402_1`. 62

R. Erdélyi and I. Ballai. Heating of the solar and stellar coronae: a review. *Astronomische Nachrichten*, 328:726–733, October 2007. doi: 10.1002/asna.200710803. 17, 18

Yue Fei, Anthony A. Chan, Scot R. Elkington, and Michael J. Wiltberger. Radial diffusion and mhd particle simulations of relativistic electron transport by ulf waves in the september 1998 storm. *Journal of Geophysical Research: Space Physics*, 111(A12):n/a–n/a, 2006. ISSN 2156-2202. doi: 10.1029/2005JA011211. URL `http://dx.doi.org/10.1029/2005JA011211`. A12209. 90

Bengt Fornberg and Natasha Flyer. *A primer on radial basis functions with applications to the geosciences.* SIAM, 2015. ISBN 978-1-61197-402-7. 93

R. H. W. Friedel, H. Korth, M. G. Henderson, M. F. Thomsen, and J. D. Scudder. Plasma sheet access to the inner magnetosphere. *Journal of Geophysical Research: Space Physics*, 106(A4):5845–5858, 2001. ISSN 2156-2202. doi: 10.1029/2000JA003011. URL `http://dx.doi.org/10.1029/2000JA003011`. 44

Roger Frigola-Alcalde. *Bayesian time series learning with Gaussian processes.* PhD thesis, University of Cambridge, 2016. 142

Carl-Gunne Fälthammar. Effects of time-dependent electric fields on geomagnetically trapped radiation. *Journal of Geophysical Research*, 70 (11):2503–2516, 1965. ISSN 2156-2202. doi: 10.1029/JZ070i011p02503. URL `http://dx.doi.org/10.1029/JZ070i011p02503`. 88

N Yu Ganushkina, TI Pulkkinen, MV Kubyshkina, HJ Singer, and CT Russell. Long-term evolution of magnetospheric current systems during storms. In *Annales Geophysicae*, volume 22, pages 1317–1334, 2004. 33

N. Yu. Ganushkina, M. W. Liemohn, M. V. Kubyshkina, R. Ilie, and H. J. Singer. Distortions of the magnetic field by storm-time current systems in earth's magnetosphere. *Annales Geophysicae*, 28(1):123–140, 2010. doi: 10.5194/angeo-28-123-2010. URL `http://www.ann-geophys.net/28/123/2010/`. 33

R. A. García, S. Turck-Chièze, S. J. Jiménez-Reyes, J. Ballot, P. L. Pallé, A. Eff-Darwich, S. Mathur, and J. Provost. Tracking Solar Gravity Modes: The Dynamics of the Solar Core. *Science*, 316:1591, June 2007. doi: 10.1126/science.1140598. 16

Paul Gaskell, Frank McGroarty, and Thanassis Tiropanis. Signal diffusion mapping: Optimal forecasting with time-varying lags. *Journal of Forecasting*, 35(1):70–85, 2015. doi: 10.1002/for.2365. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/for.2365`. 115

Mark Girolami. Orthogonal series density estimation and the kernel eigenvalue problem. *Neural Computation*, 14(3):669–688, March 2002. ISSN 0899-7667. doi: 10.1162/089976602317250942. URL `http://dx.doi.org/10.1162/089976602317250942`. 40

H Gleisner, H Lundstedt, and P Wintoft. Predicting geomagnetic storms from solar-wind data using time-delay neural networks. In *Annales Geophysicae*, volume 14, page 679. Citeseer, 1996. 62

W. D. Gonzalez, J. A. Joselyn, Y. Kamide, H. W. Kroehl, G. Rostoker, B. T. Tsurutani, and V. M. Vasyliunas. What is a geomagnetic storm? *Journal of Geophysical Research: Space Physics*, 99(A4): 5771–5792, 1994. ISSN 2156-2202. doi: 10.1029/93JA02867. URL `http://dx.doi.org/10.1029/93JA02867`. 61

Thore Graepel. Solving noisy linear operator equations by gaussian processes: Application to ordinary and partial differential equations. January 2003. URL `https://www.aaai.org/Papers/ICML/2003/ICML03-033.pdf`. 92

C. W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–438, 1969. ISSN 00129682, 14680262. URL `http://www.jstor.org/stable/1912791`. 112

Aditya Grover, Ashish Kapoor, and Eric Horvitz. A deep hybrid model for weather forecasting. In *KDD*, pages 379–386, 2015. 112

M. A. Gruet, M. Chandorkar, A. Sicard, and E. Camporeale. Multiple-hour-ahead forecast of the dst index using a combination of long short-term memory neural network and gaussian process. *Space Weather*, 16(11): 1882–1896, 2018. doi: 10.1029/2018SW001898. URL `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2018SW001898`.

Marina Gruet. *Intelligence artificielle et prévision de l'impact de l'activité solaire sur l'environnement magnétique terrestre*. PhD thesis, 2018. 60

Robin Gubby and John Evans. Space environment effects and satellite design. *Journal of Atmospheric and Solar-Terrestrial Physics*, 64(16): 1723 – 1733, 2002. ISSN 1364-6826. doi: https://doi.org/10.1016/

S1364-6826(02)00122-0. URL http://www.sciencedirect.com/science/article/pii/S1364682602001220. Space Weather Effects on Technological Systems. 88

S. Haaland, C. Munteanu, and B. Mailyan. Solar wind propagation delay: Comment on minimum variance analysis-based propagation of the solar wind observations: Application to real-time global magnetohydrodynamic simulations by A. Pulkkinen and L. Raststatter. *Space Weather*, 8(6), 2010. doi: 10.1029/2009SW000542. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2009SW000542. 112

Heikki Haario, Eero Saksman, and Johanna Tamminen. An adaptive metropolis algorithm. *Bernoulli*, 7(2):223–242, 04 2001. URL https://projecteuclid.org:443/euclid.bj/1080222083. 98

Brian Harvey. *Russian planetary exploration: history, development, legacy and prospects*. Springer Science & Business Media, 2007. 18

Simon Haykin. *Neural networks: a comprehensive foundation*. Prentice Hall PTR, 1994. 62

Philipp Hennig, Michael A Osborne, and Mark Girolami. Probabilistic numerics and uncertainty in computations. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 471(2179): 20150142, 2015. 93

Sepp Hochreiter. Untersuchungen zu dynamischen neuronalen netzen. *Diploma, Technische Universität München*, 91(1), 1991. 66

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 63

Thomas Hofmann, Bernhard Schölkopf, and Alexander J. Smola. Kernel methods in machine learning. *Ann. Statist.*, 36(3):1171–1220, 06 2008. doi: 10.1214/009053607000000677. URL http://dx.doi.org/10.1214/009053607000000677. 40, 41

T. Iyemori, H. Maeda, and T. Kamei. Impulse response of geomagnetic indices to interplanetary magnetic field. *Journal of Geomagnetism and Geoelectricity*, 31:1–9, 1979. doi: 10.5636/jgg.31.1. 62

Eun Young Ji, Y. J. Moon, N. Gopalswamy, and D. H. Lee. Comparison of Dst forecast models for intense geomagnetic storms. *Journal of Geophysical Research: Space Physics*, 117(3):1–9, 2012. ISSN 21699402. doi: 10.1029/2011JA016872. 48, 64

E.J. Kansa. Multiquadrics—a scattered data approximation scheme with applications to computational fluid-dynamics—ii solutions to parabolic, hyperbolic and elliptic partial differential equations. *Computers & Mathematics with Applications*, 19(8):147 – 161, 1990. ISSN 0898-1221. doi: https://doi.org/10.1016/0898-1221(90)90271-K. URL `http://www.sciencedirect.com/science/article/pii/089812219090271K`. 93

John G Kappenman, Lawrence J Zanetti, and William A Radasky. Geomagnetic storms can threaten electric power grid. *Earth in Space*, 9(7): 9–11, 1997. 10

William Karush. Minima of functions of several variables with inequalities as side constraints. Master's thesis, Dept. of Mathematics, Univ. of Chicago, 1939. 96

J. Kissinger, R. L. McPherron, T.-S. Hsu, and V. Angelopoulos. Steady magnetospheric convection and stream interfaces: Relationship over a solar cycle. *Journal of Geophysical Research: Space Physics*, 116(A5): n/a–n/a, 2011. ISSN 2156-2202. doi: 10.1029/2010JA015763. URL `http://dx.doi.org/10.1029/2010JA015763`. A00I19. 44

J. Kocijan. *Modelling and Control of Dynamic Systems Using Gaussian Process Models*. Advances in Industrial Control. Springer International Publishing, 2015. ISBN 9783319210216. URL `https://books.google.nl/books?id=_Gn_CgAAQBAJ`. 43

Masato Koda, Ali H Dogru, and John H Seinfeld. Sensitivity analysis of partial differential equations with application to reaction and diffusion processes. *Journal of Computational Physics*, 30(2):259 – 282, 1979. ISSN 0021-9991. doi: https://doi.org/10.1016/0021-9991(79)90103-7. URL `http://www.sciencedirect.com/science/article/pii/0021999179901037`. 151

Katharina Kormann and Elisabeth Larsson. A galerkin radial basis function method for the schrödinger equation. *SIAM Journal on Scientific Computing*, 35(6):A2832–A2855, 2013. doi: 10.1137/120893975. URL `https://doi.org/10.1137/120893975`. 93

d g Krige. *A Statistical Approach to Some Mine Valuation and Allied Problems on the Witwatersrand.* publisher not identified, 1951. URL `https://books.google.co.in/books?id=M6jASgAACAAJ`. 36

Harold W Kuhn and Albert W Tucker. Nonlinear programming, in (j. neyman, ed.) proceedings of the second berkeley symposium on mathematical statistics and probability, 1951. 96

I. E. Lagaris, A. Likas, and D. I. Fotiadis. Artificial neural networks for solving ordinary and partial differential equations. *IEEE Transactions on Neural Networks*, 9(5):987–1000, Sep 1998. ISSN 1045-9227. doi: 10.1109/72.712178. 93

Matthew Lang and Mathew J. Owens. A variational approach to data assimilation in the solar wind. *Space Weather*, 17(1):59–83, 2019. doi: 10.1029/2018SW001857. URL `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2018SW001857`. 114

J. A. Lazzús, P. Vega, P. Rojas, and I. Salfate. Forecasting the dst index using a swarm-optimized neural network. *Space Weather*, 15(8): 1068–1089, 2017. doi: 10.1002/2017SW001608. URL `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2017SW001608`. xv, 63, 65, 75

Peter M Lee. *Bayesian statistics.* Arnold Publication, 1997. 91

M. W. Liemohn, J. U. Kozyra, C. R. Clauer, and A. J. Ridley. Computational analysis of the near-earth magnetospheric current system during two-phase decay storms. *Journal of Geophysical Research: Space Physics*, 106(A12):29531–29542, 2001. ISSN 2156-2202. doi: 10.1029/2001JA000045. URL `http://dx.doi.org/10.1029/2001JA000045`. 33

Jon A Linker, Z Mikić, Doug A Biesecker, Robert J Forsyth, Sarah E Gibson, Alan J Lazarus, A Lecinski, Pete Riley, Adam Szabo, and Barbara J Thompson. Magnetohydrodynamic modeling of the solar corona during whole sun month. *Journal of Geophysical Research: Space Physics*, 104 (A5):9809–9830, 1999. 19, 114

A Loskutov, IA Istomin, KM Kuzanyan, and OL Kotlyarov. Testing and forecasting the time series of the solar activity by singular spectrum analysis. *Nonlinear Phenomena in Complex Systems*, 4(1):47–57, 2001a. 35

A Yu Loskutov, IA Istomin, OL Kotlyarov, and KM Kuzanyan. A study of the regularities in solar magnetic activity by singular spectral analysis. *Astronomy Letters*, 27(11):745–753, 2001b. 35

Richard A. Lovett. What if the biggest solar storm on record happened today? URL `https://news.nationalgeographic.com/news/2011/03/110302-solar-flares-sun-storms-earth-danger-carrington-event-science`. 8

H Lundstedt and P Wintoft. Prediction of geomagnetic storms from solar wind data with the use of a neural network. In *Annales Geophysicae*, volume 12, pages 19–24. Springer, 1994. 62

H. Lundstedt, H. Gleisner, and P. Wintoft. Operational forecasts of the geomagnetic dst index. *Geophysical Research Letters*, 29(24):34–1–34–4, 2002. ISSN 1944-8007. doi: 10.1029/2002GL016151. URL `http://dx.doi.org/10.1029/2002GL016151`. 2181. 34, 62

L. R. Lyons and M. Schulz. Access of energetic particles to storm time ring current through enhanced radial “diffusion”. *Journal of Geophysical Research: Space Physics*, 94(A5):5491–5496, 1989. ISSN 2156-2202. doi: 10.1029/JA094iA05p05491. URL `http://dx.doi.org/10.1029/JA094iA05p05491`. 89

Donald W Marquardt. An algorithm for least-squares estimation of nonlinear parameters. *Journal of the society for Industrial and Applied Mathematics*, 11(2):431–441, 1963. 69

James Clerk Maxwell. Viii. a dynamical theory of the electromagnetic field. *Philosophical transactions of the Royal Society of London*, (155):459–512, 1865. 9

J.L. McClain. *Japan, a Modern History.* W.W. Norton & Company, 2002. ISBN 9780393041569. URL `https://books.google.nl/books?id=obYhbzN-dY0C`. 6

R. L. McPherron, T. Terasawa, and A. Nishida. Solar wind triggering of substorm expansion onset. *Journal of geomagnetism and geoelectricity*, 38(11):1089–1108, 1986. doi: 10.5636/jgg.38.1089. 44

Robert L. McPherron, George Siscoe, Nancy U. Crooker, and Nick Arge. *Probabilistic Forecasting of the Dst Index*, pages 203–210. American Geophysical Union, 2013. ISBN 9781118666098. doi: 10.1029/155GM22. URL `http://dx.doi.org/10.1029/155GM22`. 35

Siamak Mehrkanoon and Johan A.K. Suykens. Ls-svm approximate solution to linear time varying descriptor systems. *Automatica*, 48(10):2502 – 2511, 2012. ISSN 0005-1098. doi: https://doi.org/10.1016/j.automatica.2012.06.095. URL `http://www.sciencedirect.com/science/article/pii/S0005109812003652`. 93

Siamak Mehrkanoon and Johan A.K. Suykens. Learning solutions to partial differential equations using ls-svm. *Neurocomputing*, 159(Supplement C): 105 – 116, 2015. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2015.02.013. URL `http://www.sciencedirect.com/science/article/pii/S0925231215001629`. 93

Siamak Mehrkanoon, Saeid Mehrkanoon, and Johan A.K. Suykens. Parameter estimation of delay differential equations: An integration-free ls-svm approach. *Communications in Nonlinear Science and Numerical Simulation*, 19(4):830 – 841, 2014. ISSN 1007-5704. doi: https://doi.org/10.1016/j.cnsns.2013.07.024. URL `http://www.sciencedirect.com/science/article/pii/S1007570413003444`. 93

Nicole Meyer-Vernet. *Basics of the solar wind.* Cambridge University Press, 2007. 18

SK Morley, JP Sullivan, MR Carver, RM Kippen, RHW Friedel, GD Reeves, and MG Henderson. Energetic particle data from the global positioning system constellation. *Space Weather*, 15(2):283–289, 2017. 63, 64

V. Munsami. Determination of the effects of substorms on the storm-time ring current using neural networks. *Journal of Geophysical Research: Space Physics*, 105(A12):27833–27840, 2000. ISSN 2156-2202. doi: 10.1029/2000JA000041. URL `http://dx.doi.org/10.1029/2000JA000041`. 33

C. Munteanu, S. Haaland, B. Mailyan, M. Echim, and K. Mursula. Propagation delay of solar wind discontinuities: Comparing different methods and evaluating the effect of wavelet denoising. *Journal of Geophysical Research: Space Physics*, 118(7):3985–3994, 2013. doi: 10.1002/jgra.50429. URL `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/jgra.50429`. 112, 124

Radford M. Neal. *Bayesian Learning for Neural Networks*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 1996. ISBN 0387947248. 36, 46

Yurii Nesterov. A method for unconstrained convex minimization problem with the rate of convergence o (1/kˆ 2). In *Doklady AN USSR*, volume 269, pages 543–547, 1983. 69

Marcia Neugebauer and Conway W Snyder. Mariner 2 observations of the solar wind: 1. average properties. *Journal of Geophysical Research*, 71 (19):4469–4484, 1966. 18

P.T. Newell, K. Liou, J.W. Gjerloev, T. Sotirelis, S. Wing, and E.J. Mitchell. Substorm probabilities are best predicted from solar wind speed. *Journal of Atmospheric and Solar-Terrestrial Physics*, 146:28 – 37, 2016. ISSN 1364-6826. doi: http://dx.doi.org/10.1016/j.jastp.2016.04.019. URL `http://www.sciencedirect.com/science/article/pii/S1364682616301195`. 44

T. Paul O'Brien and Robert L. McPherron. An empirical phase space analysis of ring current dynamics: Solar wind control of injection and decay. *Journal of Geophysical Research: Space Physics*, 105(A4):7707–7719, 2000. ISSN 2156-2202. doi: 10.1029/1998JA000437. URL `http://dx.doi.org/10.1029/1998JA000437`. 34

D. Odstrčil. Modeling 3-d solar wind structure. *Advances in Space Research*, 32(4):497 – 506, 2003. ISSN 0273-1177. doi: https://doi.org/10.1016/S0273-1177(03)00332-6. URL `http://www.sciencedirect.com/science/`

article/pii/S0273117703003326. Heliosphere at Solar Maximum. 19, 114

D. Odstrčil and V. J. Pizzo. Three-dimensional propagation of coronal mass ejections (cmes) in a structured solar wind flow: 1. cme launched within the streamer belt. *Journal of Geophysical Research: Space Physics*, 104(A1):483–492, 1999a. doi: 10.1029/1998JA900019. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/1998JA900019. 19, 114

D. Odstrčil and V. J. Pizzo. Three-dimensional propagation of coronal mass ejections (cmes) in a structured solar wind flow: 2. cme launched adjacent to the streamer belt. *Journal of Geophysical Research: Space Physics*, 104(A1):493–503, 1999b. doi: 10.1029/1998JA900038. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/1998JA900038. 19, 114

D. Odstrčil, Z. Smith, and M. Dryer. Distortion of the heliospheric plasma sheet by interplanetary shocks. *Geophysical Research Letters*, 23(18): 2521–2524, 1996. doi: 10.1029/96GL00159. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/96GL00159. 19, 114

D. Odstrčil, P. Riley, and X. P. Zhao. Numerical simulation of the 12 may 1997 interplanetary cme event. *Journal of Geophysical Research: Space Physics*, 109(A2), 2004. doi: 10.1029/2003JA010135. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2003JA010135. 19, 114

Mathew J. Owens and Robert J. Forsyth. The heliospheric magnetic field. *Living Reviews in Solar Physics*, 10(1):5, Nov 2013. ISSN 1614-4961. doi: 10.12942/lrsp-2013-5. URL https://doi.org/10.12942/lrsp-2013-5. xiv, 19, 20

Mathew J. Owens and Pete Riley. Probabilistic solar wind forecasting using large ensembles of near-sun conditions with a simple one-dimensional ”upwind” scheme. *Space Weather*, 15(11):1461–1474, 2017. doi: 10.1002/2017SW001679. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2017SW001679. 114

Mathew J Owens, Mike Lockwood, Ed Hawkins, Ilya Usoskin, Gareth S Jones, Luke Barnard, Andrew Schurer, and John Fasullo. The maunder minimum and the little ice age: an update from recent reconstructions and climate simulations. *Journal of Space Weather and Space Climate*, 7:A33, 2017. 23

G. Pallocchia, E. Amata, G. Consolini, M. F. Marcucci, and I. Bertello. Geomagnetic Dst index forecast based on IMF data only. *Annales Geophysicae*, 24(3):989–999, May 2006. URL `https://hal.archives-ouvertes.fr/hal-00318011`. 35

E. N. Parker. Cosmic-ray modulation by solar wind. *Phys. Rev.*, 110: 1445–1449, Jun 1958a. doi: 10.1103/PhysRev.110.1445. URL `https://link.aps.org/doi/10.1103/PhysRev.110.1445`. 18

E. N. Parker. The Hydrodynamic Theory of Solar Corpuscular Radiation and Stellar Winds. *The Astrophysical Journal*, 132:821, Nov 1960. doi: 10.1086/146985. 18

E. N. Parker. Dynamical Theory of the Solar Wind. *Space Science Reviews*, 4(5-6):666–708, Sep 1965. doi: 10.1007/BF00216273. 18

Eugene N Parker. Dynamics of the interplanetary gas and magnetic fields. *The Astrophysical Journal*, 128:664, 1958b. 18

John W. Pearson. A radial basis function method for solving pde-constrained optimization problems. *Numerical Algorithms*, 64(3):481–506, Nov 2013. ISSN 1572-9265. doi: 10.1007/s11075-012-9675-6. URL `https://doi.org/10.1007/s11075-012-9675-6`. 93

Diego Pennacchioli, Michele Coscia, Salvatore Rinzivillo, Fosca Giannotti, and Dino Pedreschi. The retail market as a complex system. *EPJ Data Sci.*, 3(1):33, 2014. doi: 10.1140/epjds/s13688-014-0033-x. URL `https://doi.org/10.1140/epjds/s13688-014-0033-x`. 112

Jens Pomoell and Stefaan Poedts. Euhforia: European heliospheric forecasting information asset. *Journal of Space Weather and Space Climate*, 8:A35, 2018. 19, 114

E. Priest and T. Forbes. *Magnetic Reconnection*. February 2007. 61

M. I. Pudovkin and V. S. Semenov. On the Perreault-Akasofu energy function epsilon. *Geomagnetism and Aeronomy*, 26:1026–1028, dec 1986. 44

Maziar Raissi. Deep hidden physics models: Deep learning of nonlinear partial differential equations. *The Journal of Machine Learning Research*, 19(1):932–955, 2018. 93

Maziar Raissi, Paris Perdikaris, and George Em Karniadakis. Numerical gaussian processes for time-dependent and nonlinear partial differential equations. *SIAM Journal on Scientific Computing*, 40(1):A172–A198, 2018. 92

Syama Sundar Rangapuram, Matthias W. Seeger, Jan Gasthaus, Lorenzo Stella, Yuyang Wang, and Tim Januschowski. Deep state space models for time series forecasting. In *NeurIPS 2018*, pages 7796–7805, 2018. 112

Carl Edward Rasmussen and Hannes Nickisch. Gaussian processes for machine learning (gpml) toolbox. *Journal of machine learning research*, 11(Nov):3011–3015, 2010. URL `http://www.gaussianprocess.org/gpml/code`. 71

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005. ISBN 026218253X. 36, 39, 42, 44, 45, 71, 92, 141, 145

Martin A. Reiss, Peter J. MacNeice, Leila M. Mays, Charles N. Arge, Christian Möstl, Ljubomir Nikolic, and Tanja Amerstorfer. Forecasting the ambient solar wind with numerical models. i. on the implementation of an operational framework. *The Astrophysical Journal Supplement Series*, 240(2):35, feb 2019. doi: 10.3847/1538-4365/aaf8b3. URL `https://iopscience.iop.org/article/10.3847/1538-4365/aaf8b3/pdf`. xx, 113, 132, 133

P. Riley and R. Lionello. Mapping solar wind streams from the sun to 1 au: A comparison of techniques. *Solar Physics*, 270(2):575–592, Jun 2011. ISSN 1573-093X. doi: 10.1007/s11207-011-9766-x. URL `https://doi.org/10.1007/s11207-011-9766-x`. 114, 126

Stephen Roberts, Michael Osborne, Mark Ebden, Steven Reece, Neale Gibson, and Suzanne Aigrain. Gaussian processes for time-series modelling. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371(1984):20110550, 2013. 43, 145

Juan G Roederer. *Dynamics of geomagnetically trapped radiation*, volume 2. Springer-Verlag Berlin Heidelberg, 1970. ISBN 978-3-642-49300-3. doi: 10.1007/978-3-642-49300-3. 25, 29, 88

David E Rummelhart, James L McClelland, PDP Research Group, et al. Parallel distributed processing: Explorations in the microstructure of cognition. vol. 1, 1986. 63

CT Russell. Solar wind and interplanetary magnetic field: A tutorial. *Space Weather*, pages 73–89, 2001. 17

Ali Safdari-Vaighani, Alfa Heryudono, and Elisabeth Larsson. A radial basis function partition of unity collocation method for convection–diffusion equations arising in financial applications. *Journal of Scientific Computing*, 64(2):341–367, Aug 2015. ISSN 1573-7691. doi: 10.1007/s10915-014-9935-9. URL `https://doi.org/10.1007/s10915-014-9935-9`. 93

H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1):43–49, 1978. 115

Simo Särkkä. *Linear Operators and Stochastic Partial Differential Equations in Gaussian Process Regression*, pages 151–158. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011. ISBN 978-3-642-21738-8. doi: 10.1007/978-3-642-21738-8_20. URL `https://doi.org/10.1007/978-3-642-21738-8_20`. 92

Kenneth H Schatten. Current sheet magnetic model for the solar corona. 1971. 19, 113

Kenneth H Schatten, John M Wilcox, and Norman F Ness. A model of interplanetary and coronal magnetic fields. *Solar Physics*, 6(3):442–455, 1969. 18, 113

Bernhard Scholkopf and Alexander J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, USA, 2001. ISBN 0262194759. 40, 41

Michael Schulz and Louis J Lanzerotti. *Particle diffusion in the radiation belts*, volume 7. Springer-Verlag Berlin Heidelberg, 1974. ISBN 978-3-642-65677-4. doi: 10.1007/978-3-642-65675-0. 28, 88

R. S. Selesnick, J. B. Blake, W. A. Kolasinski, and T. A. Fritz. A quiescent state of 3 to 8 mev radiation belt electrons. *Geophysical Research Letters*, 24(11):1343–1346, 1997. ISSN 1944-8007. doi: 10.1029/97GL51407. URL `http://dx.doi.org/10.1029/97GL51407`. 90

Javad Sharifi, Babak N. Araabi, and Caro Lucas. Multi-step prediction of dst index using singular spectrum analysis and locally linear neurofuzzy modeling. *Earth, Planets and Space*, 58(3):331–341, 2006. ISSN 1880-5981. doi: 10.1186/BF03351929. URL `http://dx.doi.org/10.1186/BF03351929`. 35

Xingjian Shi and Dit-Yan Yeung. Machine learning for spatiotemporal sequence forecasting: A survey. *ArXiv*, abs/1808.06865, 2018. 112

Yuri Y. Shprits, Nigel P. Meredith, and Richard M. Thorne. Parameterization of radiation belt electron loss timescales due to interactions with chorus waves. *Geophysical Research Letters*, 34(11):n/a–n/a, 2007. ISSN 1944-8007. doi: 10.1029/2006GL029050. URL `http://dx.doi.org/10.1029/2006GL029050`. L11110. 90, 99, 100

Fernando M. Silva and Luís B. Almeida. Acceleration techniques for the backpropagation algorithm. In Luis B. Almeida and Christian J. Wellekens, editors, *Neural Networks*, pages 110–119, Berlin, Heidelberg, 1990. Springer Berlin Heidelberg. ISBN 978-3-540-46939-1. 69

A. K. Singh, Devendraa Siingh, and R. P. Singh. Space weather: Physics, effects and predictability. *Surveys in Geophysics*, 31(6):581–638, Dec 2010. ISSN 1573-0956. doi: 10.1007/s10712-010-9103-1. URL `https://doi.org/10.1007/s10712-010-9103-1`. 62

John Skilling. *Bayesian Solution of Ordinary Differential Equations*, pages 23–37. Springer Netherlands, Dordrecht, 1992. ISBN 978-94-017-2219-3. doi: 10.1007/978-94-017-2219-3_2. URL https://doi.org/10.1007/978-94-017-2219-3_2. 92

E. Spencer, P. Kasturi, S. Patra, W. Horton, and M. L. Mays. Influence of solar wind–magnetosphere coupling functions on the dst index. *Journal of Geophysical Research: Space Physics*, 116(A12):n/a–n/a, 2011. ISSN 2156-2202. doi: 10.1029/2011JA016780. URL http://dx.doi.org/10.1029/2011JA016780. A12235. 44

James L. Green Sten F. Odenwald. Bracing the satellite infrastructure for a solar superstorm. URL https://www.scientificamerican.com/article/bracing-for-a-solar-superstorm. 8

M. Sugiura. Hourly values of equatorial dst for the igy. *Ann. Int. Geophys. Yr.*, 35, 1 1964. 61

T. Tao. *An Introduction to Measure Theory*. Graduate studies in mathematics. American Mathematical Society, 2011. ISBN 9780821869192. URL https://books.google.nl/books?id=HoGDAwAAQBAJ. 36

The White House, October 2016. URL https://obamawhitehouse.archives.gov/the-press-office/2016/10/13/executive-order-coordinating-efforts-prepare-nation-space-weather-events. 63

Theano Development Team. Theano: A Python framework for fast computation of mathematical expressions. *arXiv e-prints*, abs/1605.02688, May 2016. URL http://arxiv.org/abs/1605.02688. 69

Tijmen Tieleman and Geoffrey Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2):26–31, 2012. URL http://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf. 69

Martin Tillenius, Elisabeth Larsson, Erik Lehto, and Natasha Flyer. A scalable rbf–fd method for atmospheric flow. *Journal of Computational*

Physics, 298:406 – 422, 2015. ISSN 0021-9991. doi: https://doi.org/10.1016/j.jcp.2015.06.003. URL `https://www.sciencedirect.com/science/article/pii/S0021999115003824`. 93

Ioannis G. Tsoulos, Dimitris Gavrilis, and Euripidis Glavas. Solving differential equations with constructed neural networks. *Neurocomputing*, 72(10):2385 – 2391, 2009. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2008.12.004. URL `http://www.sciencedirect.com/science/article/pii/S0925231208005560`. Lattice Computing and Natural Computing (JCIS 2007) / Neural Networks in Intelligent Systems Designn (ISDA 2007). 93

Ryan Darby Turner. *Gaussian processes for state space models and change point detection*. PhD thesis, University of Cambridge, 2012. 142

James A. Van Allen and Louis A. Frank. Radiation around the earth to a radial distance of 107,400 km. *Nature*, 183(4659):430–434, 02 1959. URL `http://dx.doi.org/10.1038/183430a0`. 88

A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. J. Lang. Phoneme recognition using time-delay neural networks. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37(3):328–339, March 1989. ISSN 0096-3518. doi: 10.1109/29.21701. 62

Martin Walt. *Radial Diffusion of Trapped Particles*, pages 410–415. Springer Netherlands, Dordrecht, 1970. ISBN 978-94-010-3284-1. doi: 10.1007/978-94-010-3284-1_38. URL `https://doi.org/10.1007/978-94-010-3284-1_38`. 29

C. B. Wang, J. K. Chao, and C.-H. Lin. Influence of the solar wind dynamic pressure on the decay and injection of the ring current. *Journal of Geophysical Research: Space Physics*, 108(A9):n/a–n/a, 2003. ISSN 2156-2202. doi: 10.1029/2003JA009851. URL `http://dx.doi.org/10.1029/2003JA009851`. 1341. 34

Jack Wang, Aaron Hertzmann, and David J Fleet. Gaussian process dynamical models. In *Advances in neural information processing systems*, pages 1441–1448, 2006. 43

Jack M Wang, David J Fleet, and Aaron Hertzmann. Gaussian process dynamical models for human motion. *IEEE transactions on pattern analysis and machine intelligence*, 30(2):283–298, 2007. 43

Y.-M. Wang and N. R. Sheeley, Jr. Solar wind speed and coronal flux-tube expansion. *The Astrophysical Journal*, 355:726–732, June 1990. doi: 10.1086/168805. 35, 113

H. L. Wei, S. A. Billings, and M. Balikhin. Prediction of the dst index using multiresolution wavelet models. *Journal of Geophysical Research: Space Physics*, 109(A7):n/a–n/a, 2004. ISSN 2156-2202. doi: 10.1029/2003JA010332. URL `http://dx.doi.org/10.1029/2003JA010332`. A07212. 34

H. L. Wei, S. A. Billings, and M. A. Balikhin. Wavelet based non-parametric narx models for nonlinear input–output system identification. *International Journal of Systems Science*, 37(15):1089–1096, 2006. doi: 10.1080/00207720600903011. URL `http://dx.doi.org/10.1080/00207720600903011`. 34

D. T. Welling. The long-term effects of space weather on satellite operations. *Annales Geophysicae*, 28:1361–1367, June 2010. doi: 10.5194/angeo-28-1361-2010. 88

Wikisource. Littell's living age/volume 136/issue 1752/bright spot on saturn. wikisource, 2014. Online; accessed 11-June-2019. 6

Christopher KI Williams. Computation with infinite neural networks. *Neural Computation*, 10(5):1203–1216, 1998. 71

R. C. Willson, S. Gulkis, M. Janssen, H. S. Hudson, and G. A. Chapman. Observations of solar irradiance variability. *Science*, 211:700–702, February 1981. doi: 10.1126/science.211.4483.700. 23

S. Wing, J. R. Johnson, J. Jen, C.-I. Meng, D. G. Sibeck, K. Bechtold, J. Freeman, K. Costello, M. Balikhin, and K. Takahashi. Kp forecast models. *Journal of Geophysical Research: Space Physics*, 110(A4):n/a–n/a, 2005. ISSN 2156-2202. doi: 10.1029/2004JA010500. URL `http://dx.doi.org/10.1029/2004JA010500`. A04203. 34, 62

P Wintoft and H Lundstedt. Prediction of daily average solar wind velocity from solar magnetic field observations using hybrid intelligent systems. *Physics and Chemistry of the Earth*, 22(7-8):617–622, 1997. 114

Jian-Guo Wu and Henrik Lundstedt. Geomagnetic storm predictions from solar wind data with the use of dynamic neural networks. *Journal of Geophysical Research: Space Physics*, 102(A7):14255–14268, 1997. xv, 62, 65, 75

Samuel Xavier-De-Souza, J. A. K. Suykens, J. Vandewalle, and Désiré Bolle. Coupled simulated annealing. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 40(2):320–335, 2010. ISSN 10834419. doi: 10.1109/TSMCB.2009.2020435. 42

Xuepu Zhao and J. Todd Hoeksema. Prediction of the interplanetary magnetic field strength. *Journal of Geophysical Research: Space Physics*, 100(A1):19–33, 1995. doi: 10.1029/94JA02266. URL `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/94JA02266`. 19, 113

Wei-Xing Zhou and Didier Sornette. Non-parametric determination of real-time lag structure between two time series: The optimal thermal causal path method with applications to economic data. *Journal of Macroeconomics*, 28(1):195 – 224, 2006. 115, 116, 117, 124

D. Zhu, S. A. Billings, M. Balikhin, S. Wing, and D. Coca. Data derived continuous time model for the dst dynamics. *Geophysical Research Letters*, 33(4):n/a–n/a, 2006. ISSN 1944-8007. doi: 10.1029/2005GL025022. URL `http://dx.doi.org/10.1029/2005GL025022`. 34

D. Zhu, S. A. Billings, M. A. Balikhin, S. Wing, and H. Alleyne. Multi-input data derived dst model. *Journal of Geophysical Research: Space Physics*, 112(A6):n/a–n/a, 2007. ISSN 2156-2202. doi: 10.1029/2006JA012079. URL `http://dx.doi.org/10.1029/2006JA012079`. 34

Curriculum Vitae

Mandar Hemant Chandorkar was born on the 28th of December 1989 in Mumbai, India. After completing his secondary education in 2007 at Modern College of Arts, Science, and Commerce in Pune, India, he received a bachelors degree in manufacturing science and a masters degree in industrial engineering from the Indian Institute of Technology Kharagpur in 2012. In 2015, he received a masters degree in artificial intelligence from the Katholieke Universiteit Leuven. From 2015, he started a PhD project in the Multiscale Dynamics group at CWI Amsterdam, as a part of a collaboration between CWI and INRIA. The results of his doctoral research are presented in this thesis.

List of publications

1. M. A. Gruet, M. Chandorkar, A. Sicard, and E. Camporeale. Multiple-hour-ahead forecast of the dst index using a combination of long short-term memory neural network and gaussian process. *Space Weather*, 16(11):1882–1896, 2018. doi: 10.1029/2018SW001898. URL `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2018SW001898`

2. Mandar Chandorkar and Enrico Camporeale. Chapter 9 - probabilistic forecasting of geomagnetic indices using gaussian process models. In Enrico Camporeale, Simon Wing, and Jay R. Johnson, editors, *Machine Learning Techniques for Space Weather*, pages 237 – 258. Elsevier, 2018. ISBN 978-0-12-811788-0. doi: 10.1016/B978-0-12-811788-0.00009-3. URL `http://www.sciencedirect.com/science/article/pii/B9780128117880000093`

3. M. Chandorkar, E. Camporeale, and S. Wing. Probabilistic forecasting of the disturbance storm time index: An autoregressive gaussian process approach. *Space Weather*, 15(8):1004–1019, Aug 2017. ISSN 1542-7390. doi: 10.1002/2017SW001627. URL `https://agupubs.onlinelibrary.wiley.com/doi/full/10.1002/2017SW001627`

4. Enrico Camporeale, Yuri Shprits, Mandar Chandorkar, Alexander Drozdov, and Simon Wing. On the propagation of uncertainties in radiation belt simulations. *Space Weather*, 14(11):982–992, 2016